

Advances in neuromorphic computing: Expanding horizons for AI development through novel artificial neurons and in-sensor computing

Yubo Yang(杨玉波), Jizhe Zhao(赵吉哲), Yinjie Liu(刘胤洁), Xiayang Hua(华夏扬), Tianrui Wang(王天睿), Jiyuan Zheng(郑纪元), Zhibiao Hao(郝智彪), Bing Xiong(熊兵), Changzheng Sun(孙长征), Yanjun Han(韩彦军), Jian Wang(王健), Hongtao Li(李洪涛), Lai Wang(汪莱), and Yi Luo(罗毅)

Citation: Chin. Phys. B, 2024, 33 (**3**): 030702. DOI: 10.1088/1674-1056/ad1c58
Journal homepage: http://cpb.iphy.ac.cn; http://iopscience.iop.org/cpb

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

Disturbance observer-based fuzzy fault-tolerant control for high-speed trains with multiple disturbances

Qian-Ling Wang(王千龄), Cai-Qing Ma(马彩青), and Xue Lin(林雪)
Chin. Phys. B, 2023, 32 (**10**): 100701. DOI: 10.1088/1674-1056/acdfbe

rysformer: An attention-based graph neural network for properties prediction of crystals

Tian Wang(王田), Jiahui Chen(陈家辉), Jing Teng(滕婧), Jingang Shi(史金钢),Xinhua Zeng(曾新华), and Hichem Snoussi
Chin. Phys. B, 2023, 32 (**9**): 090703. DOI: 10.1088/1674-1056/ace247

Memristor-based multi-synaptic spiking neuron circuit for spiking neural network

Wenwu Jiang(蒋文武), Jie Li(李杰), Hongbo Liu(刘洪波), Xicong Qian(钱曦聪), Yuan Ge(葛源), Lidan Wang(王丽丹), and Shukai Duan(段书凯)
Chin. Phys. B, 2022, 31 (**4**): 040702. DOI: 10.1088/1674-1056/ac380b

Silicon-based optoelectronic synaptic devices

Lei Yin(尹蕾), Xiaodong Pi(皮孝东), Deren Yang(杨德仁)
Chin. Phys. B, 2020, 29 (**7**): 070703. DOI: 10.1088/1674-1056/ab973f

High-performance synaptic transistors for neuromorphic computing

Hai Zhong(钟海), Qin-Chao Sun(孙勤超), Guo Li(李果), Jian-Yu Du(杜剑宇), He-Yi Huang(黄河意), Er-Jia Guo(郭尔佳), Meng He(何萌), Can Wang(王灿), Guo-Zhen Yang(杨国桢), Chen Ge(葛琛), Kui-Juan Jin(金奎娟)
Chin. Phys. B, 2020, 29 (**4**): 040703. DOI: 10.1088/1674-1056/ab7806

TOPICAL REVIEW — Post-Moore era: Materials and device physics

Advances in neuromorphic computing: Expanding horizons for AI development through novel artificial neurons and in-sensor computing

Yubo Yang(杨玉波)[1,3,†], Jizhe Zhao(赵吉哲)[1,3,†], Yinjie Liu(刘胤洁)[1,3,†], Xiayang Hua(华夏扬)[1], Tianrui Wang(王天睿)[1], Jiyuan Zheng(郑纪元)[2], Zhibiao Hao(郝智彪)[1,2], Bing Xiong(熊兵)[1,2], Changzheng Sun(孙长征)[1,2], Yanjun Han(韩彦军)[1], Jian Wang(王健)[1], Hongtao Li(李洪涛)[1], Lai Wang(汪莱)[1,2,‡], and Yi Luo(罗毅)[1,2,§]

[1] *Department of Electronic Engineering, Tsinghua University, Beijing* 100084, *China*

[2] *Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing* 100084, *China*

(Received 19 September 2023; revised manuscript received 27 December 2023; accepted manuscript online 9 January 2024)

AI development has brought great success to upgrading the information age. At the same time, the large-scale artificial neural network for building AI systems is thirsty for computing power, which is barely satisfied by the conventional computing hardware. In the post-Moore era, the increase in computing power brought about by the size reduction of CMOS in very large-scale integrated circuits (VLSIC) is challenging to meet the growing demand for AI computing power. To address the issue, technical approaches like neuromorphic computing attract great attention because of their feature of breaking Von-Neumann architecture, and dealing with AI algorithms much more parallelly and energy efficiently. Inspired by the human neural network architecture, neuromorphic computing hardware is brought to life based on novel artificial neurons constructed by new materials or devices. Although it is relatively difficult to deploy a training process in the neuromorphic architecture like spiking neural network (SNN), the development in this field has incubated promising technologies like in-sensor computing, which brings new opportunities for multidisciplinary research, including the field of optoelectronic materials and devices, artificial neural networks, and microelectronics integration technology. The vision chips based on the architectures could reduce unnecessary data transfer and realize fast and energy-efficient visual cognitive processing. This paper reviews firstly the architectures and algorithms of SNN, and artificial neuron devices supporting neuromorphic computing, then the recent progress of in-sensor computing vision chips, which all will promote the development of AI.

Keywords: neuromorphic computing, spiking neural network (SNN), in-sensor computing, artificial intelligence

PACS: 07.05.Mh **DOI:** 10.1088/1674-1056/ad1c58

1. Introduction

Artificial intelligence (AI) is the general term for the science and technology which uses computers to simulate human intelligent behaviors and trains computers to learn human behaviors such as learning, judgment, and decision-making. The development of AI has brought huge economic benefits to mankind and has benefited all aspects of life, even as it has greatly promoted social development and brought social development into a new era.[1,2] Artificial neural network (ANN) has played significant role in the development of AI, which is the cornerstone of AI systems. ANN models can be divided into three developing generations depending on their computational units, according to Wolfgang Maass in 1996.[3] The first invented neural network model's computational unit is based on McCulloch–Pitts neurons, often called perceptron or threshold gate. The basic principle is that for the input Boolean signals (i.e., 0 and 1), the perceptron adds all the signals and compares them to the threshold, and then outputs a Boolean value of 0 or 1 (see Fig. 1(a)). This model is universal for computations with digital input and output, and every Boolean function can be computed by some multilayer perceptron with a single hidden layer.[3]

The second generation of neural networks introduces derivable and nonlinear activation functions (i.e., RELU or Sigmoid) based on the first generation. Besides, it transforms the original Boolean input linear superposition into a weighted sum of inputs (see Fig. 1(b)).[4] Therefore, it can not only calculate continuous values and strengthen the fitting characteristics but also be able to use the gradient descent method in backpropagation (namely, BP algorithm) to modify the weight coefficient of the network to enhance the computational accuracy of complex models further.[5] Based on the second-generation neural network model, various ANN neural network architectures and algorithms are proposed. A distinct historical landmark is the 2012 success of

[†]These authors contributed equally.

[‡]Corresponding author. E-mail: wanglai@mail.tsinghua.edu.cn

[§]Corresponding author. E-mail: luoy@tsinghua.edu.cn

© 2024 Chinese Physical Society and IOP Publishing Ltd

AlexNet in the ILSVRC image classification challenge.[6,7] AlexNet became known as a deep neural network (DNN), composed of eight sequential layers of end-to-end learning, including a sum of 60 million trainable parameters.[4] Since then, there has been a boom in deep learning in the AI domain of speech recognition,[8] object detection,[9–13] image recognition,[7,14,15] etc.

Nonetheless, it is evident that the complex hidden layers and huge trainable parameters put forward a severe test for DNN. On the one hand, it relies too much on a large amount of data. Once the model lacks enough data-driven, it would be difficult to train a suitable AI model. If the sample information is disturbed or attacked, the results of the network will be unpredictable.[14] The robustness generated by perturbation training alone cannot ensure the network's security.[16] On the other hand, the parameters of DNN have no specific physical or biological meaning. As a black box that relies entirely on data stacking, it is not explanatory, thus making many people skeptical about its authenticity. The most important thing is that every unit of the DNN is doing MAC operations almost all the time, creating a lot of power consumption. Although people are looking for new algorithms to achieve high classification accuracy and low computational power consumption, the high-power-consumption brought by the DNN architecture itself cannot be avoided.

In addition to algorithmic and device-level improvements, another way to improve accuracy and reduce power consumption is to find inspiration from brain science. The human brain does not need a lot of data input to produce efficient learning behavior (or high classification accuracy). What is more, the energy consumption and computational cost of sophisticated calculation in the human brain are relatively low. To reduce energy consumption and improve efficiency, people have further strengthened the research on biomimetics of biological neurons, and the experimental results accumulated in the past exclaim that the time of single-action potential (or spike) is often used in the biological nervous system to encode information.[3] Therefore, a neuromorphic computing network model that utilizes spike neurons as computational units, namely, the spike neural network (SNN), came into being.

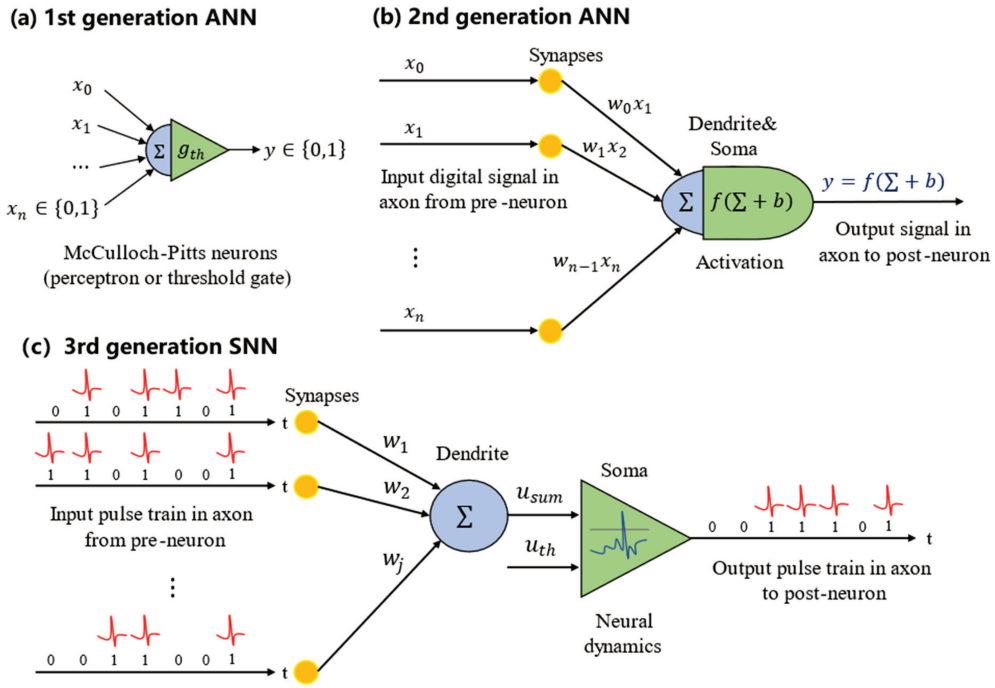


Fig. 1. Three developing generations of artificial neural networks depending on computational units.[3] (a) The first generation is based on McCulloch–Pitts neurons, namely, perceptron or threshold gate. (b) The second generation is based on weighted summation together with nonlinear activation. (c) The third generation is a more bio-explainable architecture using sequent spike inputs, integration, and comparator.

In essence, SNN belongs to a type of neuromorphic computing architecture. In addition to using neurons and synapses for computation and transmission like the second-generation networks, SNN also incorporates temporal information into the model. That is, neurons in SNN only work when the membrane potential changes (namely, when the impulse response arrives) rather than transmitting information during each propagation cycle. The typical case of SNN can be seen in Fig. 1(c). During a specific time window, the soma weights the spike pulse information arriving at each moment, and the signal in the entire time window is integrated to update the membrane potential, and when the membrane potential exceeds the threshold, the soma (neuron) will ignite to release a downstream spike.[4] The past history of each spike neuron in the time domain significantly influences the current state apart from the information propagating in the spatial environ-

ment. Since the neuron is only activated when the total membrane potential exceeds a threshold, the overall spike signal is usually sparse. In addition, MAC operations can be eliminated if the integration time window is set to 1 since the input spikes are only 0 or 1. Hence, as the third generation of neural networks, spike neural networks (SNN) have several advantages compared to traditional ANN and DNN. Compared to compute-intensive ANN networks, SNN typically requires lower power consumption because it deals with or transmits sparse data, thus improving energy efficiency. Besides, SNN utilizes pulse-based encoding, sufficiently employing temporal information of spikes to represent and process signals, which makes it suitable for time-series data processing. Last but not least, SNN mimics and emulates the working mechanism of biological neural networks, as it transmits information somewhat like synaptic transmission of electrical signals between neurons, and this merit suggests that SNN may tackle cognitive problems better than the last two generations.

In addition, over 80% of human information input in biological neural networks comes from the visual system. Individuals must continuously receive, store, and process a large amount of visual information to coordinate their activities. The retina and brain can efficiently complete this task with extremely low power consumption, relying on the exquisite multi-level structure and spike neuron mechanisms of the retinal neural networks. For example, the eyeball has different sensitivities to different targets to achieve selective reception and transmission of information, and the retina has different receptive fields to pre-process and compress information.[17,18] Completing the same visual task, the retina is four orders of magnitude higher in the energy efficiency than the traditional hardware systems. Therefore, inspired by retinal neural networks, the development of in-sensor computing is also one of the current research hotspots in the field of neuromorphic computing.

Based on the fantastic development of microelectronic devices, much progress in artificial neural/synaptic devices and circuits has been achieved.[19,20] It also has demonstrated that the fabrication of bio-mimicking circuits on a very large scale is possible.[21]

In the following parts, the SNN models mimicking the mechanism of biological neurons, together with the supporting functional materials and artificial neuron devices, are reviewed. Then, the recent progress of in-sensor computing chips, as a promising neuromorphic computing technology, is also introduced.

2. SNN

SNN architecture receives complex data, encodes them into zero–one spike sequences, and then inputs this information into SNN computational units. It is unknown how the human brain encodes the received externally rich analog information. However, people have invented several mathematical methods to process the input signals in the field of machine learning. One common approach is a frequency-based method: normalizing the original pixel intensity as the probability of emitting a pulse per unit time step, which follows a specific probability distribution, such as the Poisson distribution. If the time step is set to 20, it will produce 20 pulse matrices to get the corresponding contour, and the longer the time, the more we can reconstruct the original image. The other method uses an ANN encoder to generate a global pulse signal, which accepts the intensity value signal of multiple pixels of the picture as input, generates a pulse as an output (equivalent to feature acquisition), uses the generated pulse to do backpropagation to update the weight, and finally performs neuron calculation based on SNN architecture on the encoded input signal to realize the SNN function.

Based on the study of brain neurons, a variety of SNN mathematical models have been created, such as the spike response model (SRM), Hodgkin–Huxley (H–H) neuron model, Wilson model, Hindmarsh–Rose (H–R) model, Izhikevich spiking neuron model and leaky integrate and fired (LIF) model. Spike response model particularly describes how one neuron responds to the input spikes from other neurons. The SRM captures the temporal response properties and the effects of input spikes on the self-activation of a neuron. The core prospective of SRM is that the membrane potential of a neuron changes over time in response to inputs from other neurons, and it utilizes a linear kernel to calculate the functional relationship of membrane potential change with respect to input spikes. The mathematical expression of SRM can be expressed by the following equations:

$$V(t) = \sum_j w_j \cdot \sum_n \delta\left(t - t_j^n\right) \cdot h\left(t - t_j^n\right) + I_{\text{ext}}(t). \tag{1}$$

The H–H neuron model is a classical mathematical model that depicts the change of neuron membrane potential. The cell membrane can store electric charge and can be referred to as a capacitor C_m in circuits, while Na^+ and K^+ ion channel is modeled as a variable conductivity g in series with a battery E. E represents the equilibrium potential of a particular ion, and g reflects the permeability of the channel to the specific ion. The dynamic conductivity g depends on the voltage V_m across the membrane. Generally, a H–H model requires four ordinary differential equations as follows:

$$\begin{aligned} I &= C_m \frac{\mathrm{d}V_m}{\mathrm{d}t} + g_{\mathrm{K}^+} n^4 \left(V_m - V_{\mathrm{K}^+}\right) \\ &\quad + g_{\mathrm{Na}^+} m^3 h \left(V_m - V_{\mathrm{Na}^+}\right) + g_{\mathrm{l}} \left(V_m - V_{\mathrm{l}}\right), \end{aligned} \tag{2}$$

$$\frac{\mathrm{d}n}{\mathrm{d}t} = \alpha_n(V_m)(1-n) - \beta_n(V_m) n, \tag{3}$$

$$\frac{\mathrm{d}m}{\mathrm{d}t} = \alpha_m(V_m)(1-m) - \beta_m(V_m) m, \tag{4}$$

$$\frac{\mathrm{d}h}{\mathrm{d}t} = \alpha_h(V_m)(1-h) - \beta_h(V_m)h, \tag{5}$$

where $\alpha_i(t)$ indicates the conversion rate from prohibition to permission and $\beta_i(t)$ vice versa, and n, m, h limit the transmission of Na^+ and K^+, and the parameters with subscript l represent those of the leaking channel. The mathematical expression of SRM model is relatively simple and can obtain the temporal correlations between neurons. However, it neglects the dynamic properties of neurons and cannot fit well with complicated networks. As H–H model considers more complex situations and it needs four ordinary differential equations, it is able to capture the accurate biological characteristics of membrane potential. Nonetheless, it is rather difficult to carry out large-scale simulations. The most commonly used SNN model is the leaky integrate and fired model,[22] whose main idea is to launch a spike when the sum of collected signals (or the membrane potential) exceeds the threshold and to return to the resting potential after ignition. It also adopts a refractory period model after the ignition, which means the neuron will not be excited for a period.

It is vital for us to get a view of several partitions of a neuron and their functions. (1) The axon generates an action potential, releases neurotransmitters, and produces a pulse. (2) Dendrites or cell bodies receive neurotransmitters, generating a potential difference between inner and outer membranes to form an electrical current. (3) Leakage current is a biological concept, which is the loss of membrane potential during axonal transmission and is replaced by resistance in hardware circuits.

Neuronal dynamics can be conceived as a summation process combined with a mechanism that triggers an action potential when the sum of inputs exceeds a threshold voltage.[23] When $u_{\mathrm{res}} < u_{\mathrm{th}}$ and the summation voltage $u_i(t)$ reaches the threshold u_{th}, it will trigger action potential pulses. All the emitted pulses' shape is similar, and the transmitted information essentially lies in whether a pulse is at a specific time. This process is expressed by an integrate and fire (namely, IF) model, and its key points include (1) the formula of membrane potential, $u_i(t)$ and (2) generation mechanism of a pulse. There is only one capacitor in the IF model

$$\frac{\mathrm{d}V}{\mathrm{d}t} = \frac{1}{C}I(t). \tag{6}$$

In general, the action potential of neurons is mainly realized by the dynamic selectivity of Na^+ and K^+. The classical Hodgkin and Huxley[22] model equates the potential of various ions to a battery, refers to gated channels as a regulated variable sodium resistance, and equates the passive nature of the membrane potential to a capacitor. The above are then connected in parallel to obtain a model of the electrical properties of the neuron. If considering only the passive characteristics of membrane potential and the triggering of action potential, we can derive a LIF model that simplifies the original H–H model to a parallel connection of a capacitor, a battery, and a variable resistance, namely, leaky integrate and fire model (see Fig. 2(a)).

Usually, the LIF model is very similar to the real biological neuron. The cell membrane is not a perfect insulator (it is a capacitor), so it will constantly exchange ions inside and outside the membrane. As a consequence, the surface charge will slowly leak from the membrane (see Figs. 2(a) and 2(b)).[23] When there is only one input, the voltage will automatically leak and return to the rest state u_{res}. In the LIF model, it is generally believed that the potential will first drop below u_{res} and then rise to the resting potential (see Figs. 2(c) and 2(d)).

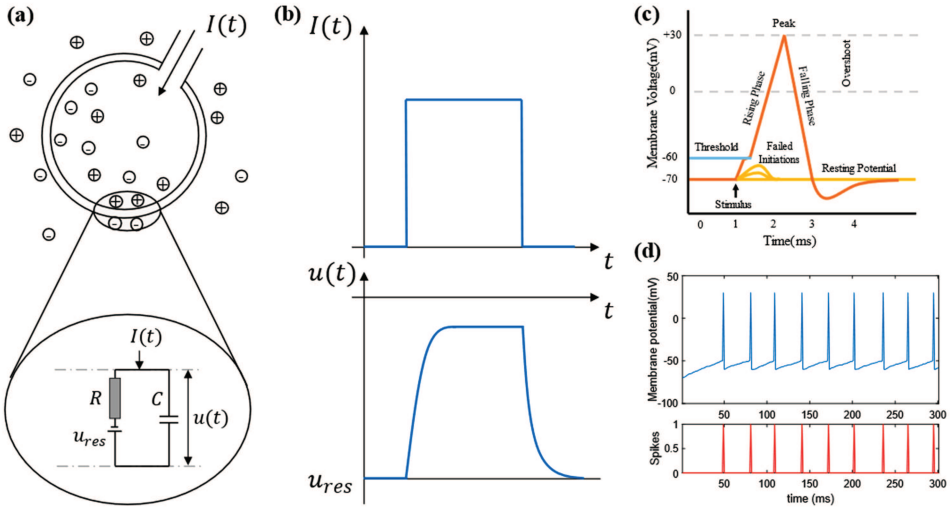


Fig. 2. Leaky integrate and fire model. (a) A neuron, which is enclosed by the cell membrane (big circle), receives a (positive) input current $I(t)$, which increases the electrical charge inside the cell. The cell membrane acts like a capacitor in parallel with a resistor, which is in line with a battery of potential u_{res} (zoomed inset). (b) The cell membrane reacts to a step current (top) with a smooth voltage trace (bottom). (c) Detailed activity of membrane potential during one pulse arrival. (d) Simulation of LIF model with $u_{\mathrm{res}} = -60$ mV, $u_{\mathrm{peak}} = 30$ mV and $u_{\mathrm{ini}} = -70$ mV.

The primary LIF circuit comprises a capacitor C and a resistor driven by current I. The whole circuit can be classified into two steps: 1) When the driving current I disappears, the voltage across the capacitor is provided by $u_{\rm res}$. The total current in the circuit is $I(t) = I_{\rm R} + I_{\rm C}$. 2) The second part is to charge the capacitor, and the current flowing through the capacitor satisfies $I_{\rm R} = \frac{{\rm d}u}{{\rm d}t}C$, so the current satisfies

$$I(t) = \frac{u(t) - u_{\rm res}}{R} + C\frac{{\rm d}u(t)}{{\rm d}t}. \tag{7}$$

By introducing the time constant $\tau_{\rm m} = RC$, we obtain (see Figs. 2(c) and 2(d))

$$\tau_{\rm m}\frac{{\rm d}u(t)}{{\rm d}t} = -\left(u(t) - u_{\rm res}\right) + RI(t). \tag{8}$$

Assuming that a spike signal is received at time 0, hence the membrane potential is $U_{\rm res} + \Delta u$. If the subsequent input current is 0, after sufficient time, the membrane potential will relax to the resting potential exponentially, and the solution of the equation is

$$u(t) - u_{\rm res} = \Delta u \exp\left(-\frac{t - t_0}{\tau_{\rm m}}\right). \tag{9}$$

If the subsequent input exists when the initial voltage is $u_{\rm res}$, the solution should be

$$u(t) = -u_{\rm res} + RI_0\Delta u\left[1 - \exp\left(-\frac{t}{\tau_{\rm m}}\right)\right]. \tag{10}$$

To summarize, the core idea of the LIF model is that if the input of the pre-neuron does not exceed the threshold, the acquired potential will return to the resting state. Suppose a neuron receives several pulses and integrates them to reach a threshold point. In that case, it will emit a pulse signal that stimulates another post-neuron, causing a refractory period after firing. Even if a stimulus is given to the firing neuron, it will not provide a corresponding response.[24] Based on the study of brain neurons, a variety of SNN mathematical models have been created, and several SNN functions have been realized using different devices, some of which will be described below.

2.1. LTP/LTD

2.1.1. Physiological mechanisms

Long-term potentiation and long-term depression are two significant phenomena (see Fig. 3(a)) in neuroscience that underlie synaptic plasticity, which correspond to the persistent strengthening or weakening of synapses due to recent activity patterns. Long-term potential (LTP) was first discovered in the rabbit hippocampus. Suppose that there are two connected neurons CA3 and CA1 (see illustration in Fig. 3(b)). When we impose a single spike on the axon of CA3, we shall obtain a low degree of potential change from B's postsynaptic membrane. If we stimulate the pre-neuron A with a high-frequency signal, we can get a high level of potential difference (2–3 times the potentials before) after imposing the same single spike on the axon of the pre-neuron. Interestingly, the phenomenon lasts for several hours or even a few days, and we refer to the appearance as long-term potentiation.[25] Regarding long-term depression, we should utilize a low-frequency stimulation on the pre-neuron instead and get a lower degree potential than the original one.

LTP is usually classified into two stages, namely, early-LTP[26] and late-LTP,[27] both of which exclaim that exerting repetitive stimulations on the presynaptic cell can enhance the synaptic efficacy of postsynaptic. The physical mechanisms of LTP and LTD are still for discovery, however, some successful mathematical models have been developed, such as the Hebbian learning model.[28,29] The core of the Hebbian learning model considers that continuous and repetitive reflexes tend to cause persistent changes in cells, increasing the cell's stability. When CA3's axons are close enough to stimulate CA1 and repeatedly to stimulate it, one or two cells will experience some metabolic changes, which makes the efficiency of CA3 (as firing cells of neuron CA1) increase.

2.1.2. Mathematical model

Hebbian theory focuses on how neurons might connect themselves to become engrams. From the perspective of ANN, it can be described as a method to determine how to change the weight between different neurons. The weight increases if both neurons are activated simultaneously and reduce if activated separately. Within binary neurons (spiking activations), connections would be set to 1 if the connected neurons have the same activation for a pattern. When several training patterns are used, the expression becomes an average of individual ones[30]

$$w_{ij} = \frac{1}{p}\sum_{k=1}^{p} x_i^k x_j^k = \langle x_i x_j \rangle, \tag{11}$$

where w_{ij} is the weight of the connection between neurons i and j, x_i^k the kth input from neuron i, and p the total number of input patterns. A classical mathematical model is BCM theory, which modifies the traditional prior-Hebbian model with biological and experimental justification.

The BCM model adopts a sliding threshold (see Fig. 1(c)) for LTP and LTD induction, and points out that a dynamic adaptation of the time-averaged postsynaptic activity stabilizes synaptic plasticity. In the BCM model, a postsynaptic neuron tends to experience the LTP process if it is in a higher-reactive state (e.g., pre-neuron is firing at high frequency) or LTD if it is in a lower-reactive condition.[31] The BCM model can be expressed as follows:

$$y = \sum_i w_i x_i, \tag{12}$$

$$\frac{\mathrm{d}w_i}{\mathrm{d}t} = y(y-\theta_y)x_i - \varepsilon w_i, \tag{13}$$

$$\theta_{\mathrm{th}} = E^{\mathrm{p}}\left[(y/y_0)\right]. \tag{14}$$

In the above equations, the neuron output is assumed to be the result of the linear superposition, and the weight decays exponentially with the magnitude of ε. The expressions also define the modified activity threshold as the power of the mathematic expectation of the neuron's output divided by a constant y_0 (see Fig. 3(d)).

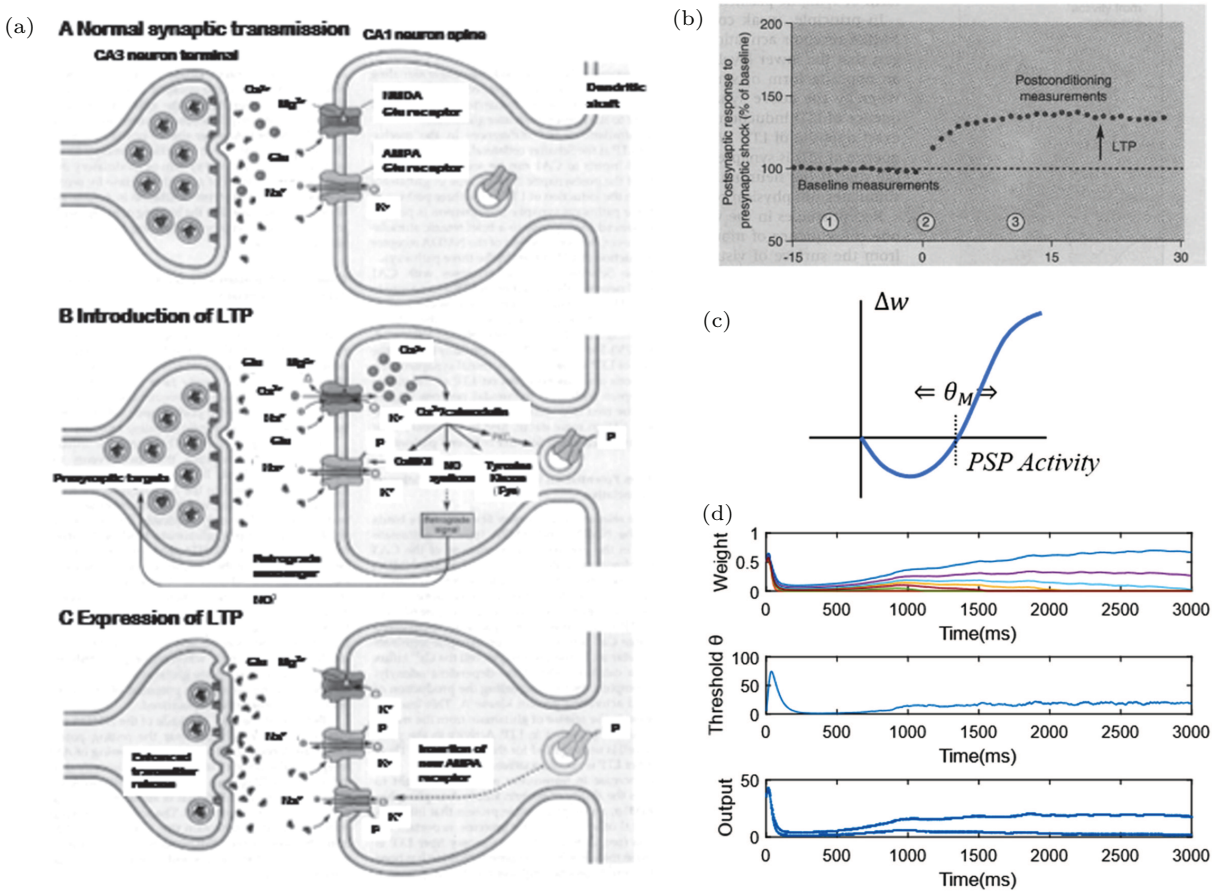


Fig. 3. Long-term potentiation and long-term depression. (a) Physiological mechanism schematic diagram of the LTP. A. During normal, low-frequency synaptic transmission, glutamate released from CA3 Schaffer collateral axons acts on both NMDA and AMPA. But sodium and K^+ can only flow through AMPA, for NMPA is blocked by Mg^{2+}. B. During high-frequency tetanus, the large depolarization of the postsynaptic membrane relieves the Mg^{2+}, allowing Ca^{2+}, Na^+, and K^+ to flow through these channels. C. Second-messenger cascades activated during induction of LTP have two main effects on synaptic transmission: phosphorylation through activation of AMPA. In addition, the postsynaptic cell releases retrograde messengers that activate protein kinases to enhance subsequent transmitter release.[32] (b) LTP is the persistent increase in synaptic strength following high-frequency stimulation of a chemical synapse.[33] (c) Sliding threshold for LTP or LTD in BCM rule. (d) Simulation of weight, threshold, and output of a-7-input-neuron model after 3000 iterations based on BCM method.

2.2. STD/STF

2.2.1. Physiological mechanisms

Synaptic plasticity is the lasting phenomenon or change of synapses' strength, function, and efficiency characteristics. Short-term memory plasticity (STP), also called dynamical synapses, refers to the appearance in which synaptic efficacy enhances or weakens over time to reflect the history of presynaptic activity.[34,35] There are two opposite effects on synaptic efficacy: Short-term depression (STD) and short-term facilitation (STF). Unlike LTP, STP has shorter time scales that last for a hundred milliseconds and will quickly return to the ground state level.[36]

The signaling process at the terminal of a presynaptic neuron's axon consumes neurotransmitters. STD should happen as the number of neurotransmitters between two short interval pulses is too poor to transmit potential signals, which leads to the decrease of the membrane potential of the postsynaptic neuron. At the same time, STF is induced by an influx of calcium ions into the axon terminal after the arrival of a spike, which raises the release probability of neurotransmitters.

STP plays a profound role in brain functions. From the perspective of computation, the time scale of STP lies between fast neural signaling (on the order of milliseconds) and experience-induced learning (on the demand of minutes or more), which has the same time scale of many behaviors in daily life such as motor control, speech recognition, and working memory. Hence, it is vital to establish a neural substrate using STP to cope with temporal information on the rel-

evant time scales. STP claims that the postsynaptic neuron response depends on the presynaptic activity history, which is a process of extracting and using historical information. In a large network, the STP can greatly enrich the network's dynamic behavior, conferring information processing capacities that would be difficult to implement using static connections on the neural system.

2.2.2. Mathematical model

According to Tsodyks and Markram's theory, STD is modeled by a normalized variable x $(0 \leq x \leq 1)$, representing the fraction of resources that remain available after neurotransmitter depletion. The STF effect is expressed by a parameter u , which denotes the fraction of available resources ready for utilization (namely, release probability). The arrival of each spike will induce the influx of calcium on the pre-synaptic membrane, which increases the release probability u. A total neurotransmitter of xu is consumed in order to produce a postsynaptic spike. Between spikes, the releasing probability u will decay back to 0 with time time-constant τ_f and x will recover to 1 with time constant τ_d.[37] To summarize, the dynamics of STP can be expressed by

$$\frac{\mathrm{d}u}{\mathrm{d}t} = -\frac{u}{\tau_f} + U\left(1-u^-\right)\delta\left(t-t_{\mathrm{sp}}\right), \tag{15}$$

$$\frac{\mathrm{d}x}{\mathrm{d}t} = -\frac{1-x}{\tau_d} - u^+x^-\delta\left(t-t_{\mathrm{sp}}\right), \tag{16}$$

$$\frac{\mathrm{d}I}{\mathrm{d}t} = -\frac{I}{\tau_d} + Au^+x^-\delta\left(t-t_{\mathrm{sp}}\right), \tag{17}$$

where t_{sp} denotes the arrival time of spikes, and U is the increment of u induced by a spike. The upper label '$\pm$' corresponds to the time after and before the arrival of spikes. It can be deduced that $u^+ = u^- + U(1-u^-)$ from Eq. (15) and $\Delta I\left(t_{\mathrm{sp}}\right) = Au^+x^-$ from Eq. (17), where A represents the response amplitude produced by the total release of all the neurotransmitters $(u = x = 1)$, called the absolute synaptic efficacy of the connections.

When $t = t_{\mathrm{sp}}$, the release probability u will rise and lead to STF. The moment STF appears, the number of available neurotransmitters should decrease, resulting in STD (see Fig. 4(a)). After the spike, u recovers to 0 with a constant τ_f and x returns to 1 with τ_d. It can be seen that u and x are inversely kinetic processes, and the product $E = ux$ characterizes whether the synaptic efficacy is dominated by STD or STF. (1) If $\tau_d \gg \tau_f$, and U is within an extensive enough parameter range, the significant drop in x caused by an initial spike takes a long time to recover to 1, and is therefore STD-dominated (see Fig. 4(b)). When $\tau_f \gg \tau_d$, and U is small, the synaptic efficacy increases gradually as the spike arrives, so the synapse is STF-dominated (see Fig. 4(c)).

STP can change the activity or number of AMPAR or NMDAR receptors on the postsynaptic membrane according to the history of presynaptic activity and then affect the distribution of potentials on membrane. STD-dominated synapses are suitable for low-frequency signal transmission, whereas STF is used for high ones. By adding a perturbation term to the firing rate, the filtering characteristics of dynamic synapses can be derived when the firing rate of the presynaptic population changes arbitrarily with time so as to improve the transmission of neuronal information.

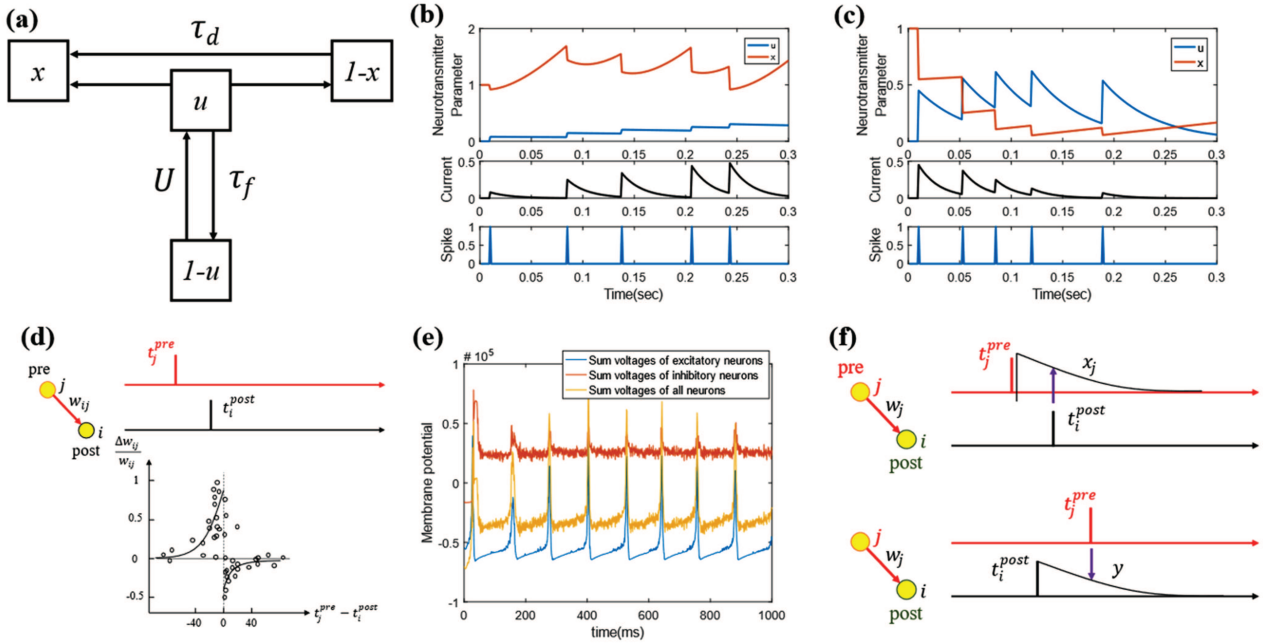


Fig. 4. Short-term plasticity and spike-timing dependent plasticity. (a) Two inversely kinetic processes of STD and STF.[37] (b) The dynamics of an STF-dominating synapse. (c) The dynamics of an STD-dominating synapse. (d) A basic model of STDP and an illustration of STDP learning window. (e) A simulation of the basic STDP model, where there are 800 excitatory neurons and 200 inhibitory neurons. (f) An illustration of online implementation of STDP models. Top: A presynaptic spike leaves a trace $x_j(t)$ which is read out (arrow) at the moment of the postsynaptic spike. The weight change is proportional to that value $x_j(t_i^{\mathrm{post}})$. Bottom: A postsynaptic spike leaves a trace $y(t)$ which is read out (arrow) at the moment of a presynaptic spike.[38]

2.3. STDP

2.3.1. Physiological mechanisms

Spike-timing dependent plasticity (STDP) is one form of Hebbian learning model with temporal asymmetry, which is influenced by the tight temporal correlations between the spikes of pre-and postsynaptic neurons. For STDP, if repeated and continuous presynaptic pulses arrive a few milliseconds before the postsynaptic pulse fires, it should undergo an LTP process. While LTD is triggered as a repetitive presynaptic pulse arrives after a postsynaptic one. The change in synaptic potential can be plotted as a function of the time difference between the pre- and post-synapses, and we call it the STDP function or the learning window (see Fig. 4(d)). This function varies depending on the type of synapse. The fast variation of the STDP function with the relative timing of spikes indicates the possibility of time coding on a millisecond time scale.

Postsynaptic neurons will fire shortly before or after injecting short current pulses to stimulate presynaptic neurons (or presynaptic pathways). The short interval pulses are typically repeated 50–100 times in pairs at a fixed frequency. The synaptic weight is measured as the postsynaptic potential's amplitude (or initial slope). The change in synaptic weight is plotted as a function of the relative time between the arrival of the presynaptic spike and the postsynaptic one (see Fig. 4(d)). STDP is considered biologically plausible. In the intact brain, action potentials are usually timed precisely according to external stimuli, although this does not apply to all brain regions and cell types.[39,40]

2.3.2. Mathematical model

As to a presynaptic neuron j, its change in weight w_j depends on the difference between presynaptic spike arrival time and postsynaptic spike time. Suppose that a fth pulse arrives at a time t_j^f and a nth pulse at t_i^n, the total weight change induced by a pair of synapses before and after, namely, the primary STDP method, can be usually defined by an equation as follows:[41,42]

$$\Delta w_j = \sum_{f=1}^{N}\sum_{n=1}^{N} W\left(t_i^n - t_j^f\right), \tag{18}$$

where function $W(X)$ represents a STDP function, or a STDP learning window. Function $W(X)$ has a more popular and typical expression

$$W(x) = \begin{cases} A_+ \exp\left(-x/\tau_{\text{pre}}\right), & x > 0, \\ -A_- \exp\left(x/\tau_{\text{post}}\right), & x < 0. \end{cases} \tag{19}$$

Parameter A_+ and A_- might rely on the values of current weight,[39] and the time constant τ is about 10 milliseconds (see Fig. 4(e)).

STDP function in the equation above can be implemented based on an online update principle by utilizing the following assumptions. Each presynaptic spike arrival leaves a trace $x_j(t)$ which is updated by an amount $a_+(x_j)$ at the moment of spike arrival and decays exponentially without spikes. Similarly, each postsynaptic spike leaves a trace y. The trace can characterize back propagation caused by potential synaptic voltage or calcium influx caused by back propagation potential.

$$\tau_+ \frac{\mathrm{d}x_j}{\mathrm{d}t} = -x_j + a_+(x_j)\sum_f \delta\left(t - t_j^f\right), \tag{20}$$

$$\tau_- \frac{\mathrm{d}y}{\mathrm{d}t} = -y + a_-(y)\sum_n \delta\left(t - t^n\right). \tag{21}$$

Consequently, the change of weight can be expressed as

$$\begin{aligned} \frac{\mathrm{d}w_j}{\mathrm{d}t} &= A_+(w_j)\,x(t)\sum_n \delta\left(t - t^n\right) \\ &\quad - A_-(w_j)\,y(t)\sum_f \delta\left(t - t_j^f\right). \end{aligned} \tag{22}$$

Therefore, weight increases to a certain number at the instant of the emission of postsynaptic pulse, and the increase depends on the presynaptic pulse trace x_j. By the same token, the weight is depressed at the moment of presynaptic spikes, and the decrease depends on the previous postsynaptic pulse trace y (see Fig. 4(f)).

The training process, weight assignment of connecting presynaptic and postsynaptic neurons, is essential for constructing SNN based on the synaptic devices. The training methods of SNN can be divided into the following types: supervised learning, unsupervised learning, and ANN transformation into SNN models. (1) Supervised learning: The typical SNN model is composed of three processes: charging, discharging, and resetting potential, and when the membrane potential exceeds the threshold, the pulse is released, which corresponds to an underivable step function, so the BP algorithm cannot be used to correct the weight. To achieve supervised learning, people use the sigmoid function to approximate the step function to realize backpropagation. Another path is to use the time-based learning method, that is, using a continuous function to substitute the discrete relationship of potential concerning time and then to derive time according to a whole series of spike signals to complete backpropagation.[43,44] These two paths are based on gradient settings and lack biological rationality. (2) Unsupervised learning: The algorithm structure is designed by imitating the synaptic plasticity mechanism existing in biological neurons, and unsupervised learning can be realized in SNN.[45,46] Synaptic plasticity refers to the impulse response of a pair of neurons in the presynaptic membrane and postsynaptic membrane that affects the weight of synapses between the two (such as STDP), and algorithms based on synaptic plasticity usually use the learning rate A and the time constant τ to characterize the potential information $u(t)$. (3) Transformation from ANN to SNN: Cao *et al.* propose a method to tailor the CNN architecture to meet the

requirement of SNN, train the untrimmed CNN in the same way as a traditional CNN would do, and finally apply the outcome weights derived from the tailored CNN to an SNN architecture.[47] In addition, there are a variety of ANN-to-SNN conversion methods.[48,49] However, the converted SNNs perform poorly during inference, making them undesirable.

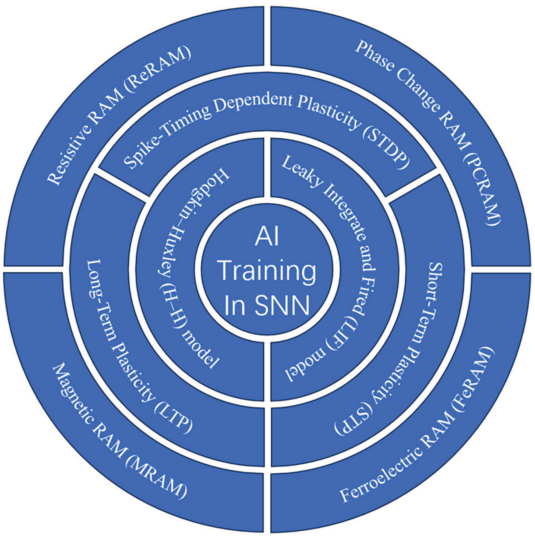


Fig. 5. The hardware implementation of SNN architecture.

The architecture of SNN is shown in Fig. 5. The LIF model, H–H model, and other models were proposed as inspired by brain neural networks to achieve neuromorphic computing with high efficiency and low energy consumption. Implementing neuromorphic computing needs neuromorphic devices that mimic synaptic plasticity (such as long-term, short-term, and spike-time-dependence plasticity). The neuromorphic devices will be discussed in the next section.

3. Artificial synapse and neuron device

For the realization of hardware-based SNNs, development of an artificial synapse and neuron is crucial. As shown in Fig. 6(a), the biological synapse is the connection between two neurons, which can memorize the synaptic weight and control the weight with synaptic plasticity. Normally, the synaptic weight is represented as a tunable electrical conductance, which can reflect the synaptic plasticity. Most of the artificial synapses accept the voltage input from the presynaptic neurons and emit the current output to the postsynaptic neurons, as a biological synapse does. So, the basic requirement of synaptic devices is that the connection strength that usually appears as conductance or resistance can be modulated by external stimulus. Since neuromorphic systems possess the ability to remember and calculate at the same time, most of the artificial synapse devices are evolved from memory devices or memory cells. Based on different physical mechanisms, resistive random access memory (ReRAM), phase change random access memory (PCRAM), ferroelectric random access memory (FeRAM), and magnetic random access memory (MRAM) are among the most promising artificial synapse devices.

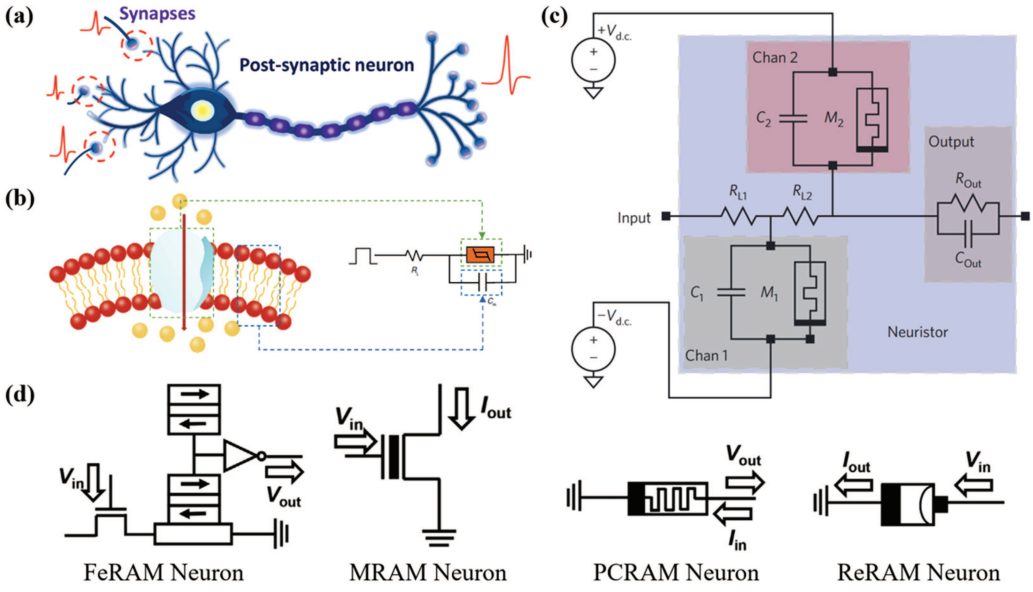


Fig. 6. (a) Biological neurons receive input from other neurons through connected synapses.[52] (b) An artificial LIF neuron model with accumulating inputs.[24] (c) A H–H neuron circuits using memristors.[51] (d) Schematic illustration of various artificial neuron devices.[52]

By contrast, the neuron generates spike signals as the output only when the input transmitted from the synapses is above a threshold as shown in Fig. 6(a). A threshold function is to judge whether signals are sufficiently collected from the synapses. Most of the artificial neurons take the current input from the previous synapses and produce the voltage output to the next synapses as the form of spiking. ReRAM, PCRAM, FeRAM and MRAM all can switch freely between high and low resistance states, which can help to realize the threshold judgement. But it is difficult to realize the neuron function in

single device. Additional devices or circuits, such as capacitor or reset circuits, are need to complete the entire neuron function. For example, the threshold switching memristor (TSM) neuron acts as a post-neuron to integrate the signals from multiple inputs via a capacitor.[50] In Fig. 6(b), when the accumulated capacitor potential approaches a certain value, the neuron fires and outputs a spiking pulse.[24] The H–H neuron model was built using memristors as shown in Fig. 6(c). This circuit can replicate complex neuronal dynamics, such as all-or-nothing pulses, bifurcation thresholds, signal gain, and the refractory period of continuous pulse firing states.[51] As illustrated in Fig. 6(d), different from voltage input (V_{in}) and current output (I_{out}) operation mode of the synapse, each artificial neuron has a different operation scheme according to whether its input and output are the voltage and current or vice versa, as the working principles differ.[52] Considering that a biological neuron accepts ions from the synapses and generates a membrane potential as the form of spiking, an artificial neuron device does better to follow this operation mode, it receives current input (I_{in}) and fires voltage output (V_{out}). It is worth noting that numerous traditional synapses update their weights according to each analog conductance change and produce I_{out}. The collected I_{out} from the synapses is applied to a neuron as I_{in}, then the neuron gives rise to V_{out}, which will be V_{in} to the next synapses. On the contrary, when the neuron device accepts V_{in} or generates I_{out}, an additional current–voltage (I–V) converter is necessary between the neuron and the synapses owing to the reversed operation mode.

3.1. ReRAM

ReRAM devices, also called memristors, usually own a simple metal/insulator/metal sandwich structure with two terminations (anode and cathode). The insulator layer between the anode and cathode is a resistive layer. When applying appropriate voltage (consecutive sweep or pulse) to the cathode, the conductivity of the resistive layer could change significantly. The conductive change can be maintained even if the voltage signal is removed, called a nonvolatile device. In another case, the resistance changes at above threshold voltage and recovers to the initial value after the voltage signal is removed, called a volatile device or threshold switching. Using this threshold switching, neuronal IF operation can be achieved. First, when the input is applied in the form of current, the charges are integrated in the membrane capacitor because the device is in the high resistance state (HRS) initially. The voltage applied to the threshold switching memristor is then elevated by the integrated charges in the membrane capacitor. When the applied voltage reaches the threshold voltage, the integrated charges are fired because the device changes to the low resistance state (LRS). This type of LIF operation induces neuronal spiking.[24]

The "set" operation is to change the device from HRS to LRS, and the "reset" is the opposite operation. As shown in Fig. 7(a), there are two types of switching characteristics: unipolar switching and bipolar. For unipolar switching, the set and reset operations can be done with voltage pulses with the same polarity, while for bipolar switching, the voltage pulses to induce the set and reset operations are of different polarities.

The mechanism of resistive change is related to the forming and rupture of filamentary in the resistive layer accompanied by the mitigation of cation or anion. The mechanism can be categorized into valence change mechanism (VCM), electrochemical metallization (ECM), and thermochemical mechanism (TCM). As TCM cannot be effective independently, subsequent discussion will be divided into VCM, ECM, and hybrid mechanisms.

3.1.1. VCM

Typically, the resistive layer material is a metal oxide film, such as HfO_2, Ta_2O_5, TiO_2, etc. Oxygen atoms could be knocked out of the lattice of metal oxide and migrate to an anode under an electric field induced by the applied voltage. Oxygen anion migration will leave oxygen vacancies, which can be viewed as a type of defect. The localized oxygen vacancies lead to the formation of the sub-oxide phase as a conductive filament, which connects the anode and cathode and results in the resistive switching from HRS to LRS. As shown in Fig. 7(b), the conductive filament system in an HfO_2-based device is a quasi-core–shell structure consisting of a metallic hexagonal-Hf_6O phase and its crystalline surroundings (monoclinic or tetragonal HfO_x). The metallic hexagonal-Hf_6O phase is a typical sub-oxide phase.[53]

To form the conductive filament, enough oxygen ions and oxygen vacancies must be imported into the resistive layer. But the pristine resistive layer, especially the devices fabricated by atom layer deposition (ALD), possesses not so many defects.[56,57] So, the applied voltage of the first-time resistive switching is usually larger than the subsequent switching processes, which guarantees the migration of enough oxygen ions to form the filament. The switching process from the original state to LRS is called forming, while subsequent switching from HRS to LRS is "set" and opposite switching is "reset". It is not desirable to own a large forming voltage in a practical device. Thus, a lot of efforts were made to fabricate the "forming-free" devices. The forming voltage is demonstrated to be linearly dependent on the thickness of the oxide film.[58] Hence, a thinner film, such as 3 nm, was designed to reduce the forming voltage. In addition, introducing oxygen vacancies in the oxide film by controlling the annealing ambient was also reported to reduce the forming voltage.[59,60]

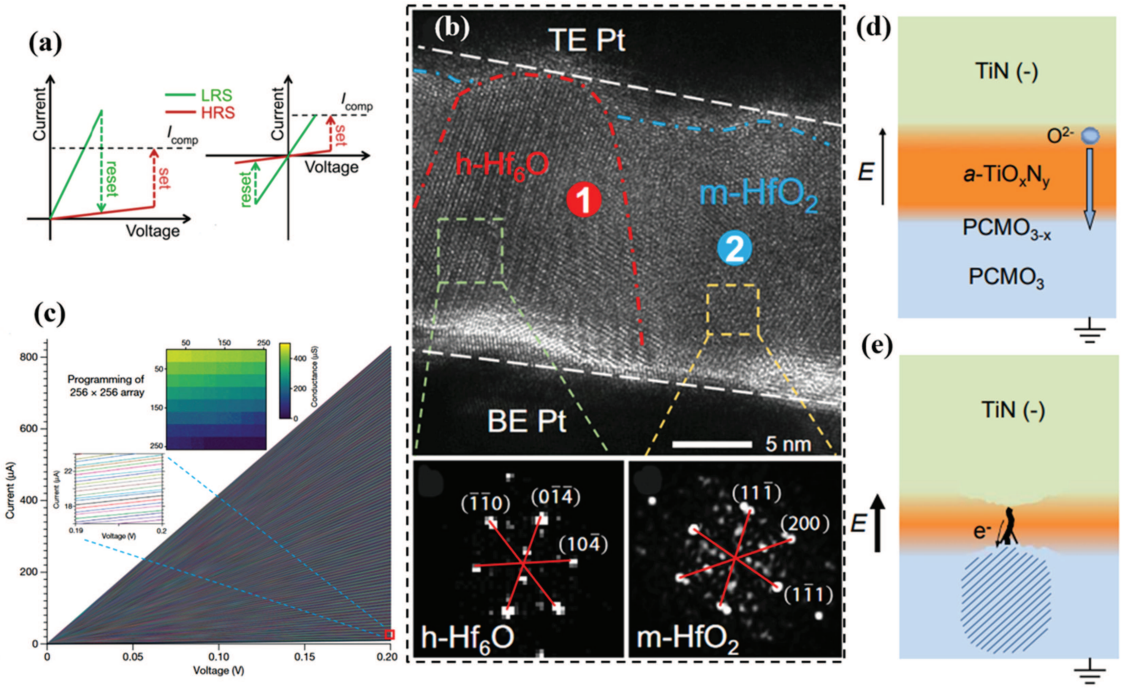


Fig. 7. VCM devices: (a) typical switching behavior of RRAM; (b) HRTEM of a complete CF in the LRS device;[53] (c) 2048 resistance levels of an individual device;[54] (d), (e) schematic models illustrating resistive switching to LRS.[55]

The reset process, switching from the LRS to the HRS, can be explained by the rupture of the conductive filament. On the contrary to the forming or set process, it is necessary to apply a negative voltage pulse or sweep on the anode, which would drive ions to migrate to the cathode and neutralize the oxygen vacancies. The sub-oxide phase would partially oxidize, while the conductive filament would partially "dissolve" and break.[61] Then the resistance state of the device will increase. However, the resistance of the device cannot return to its initial state as the conductive filament cannot be fully dissolved.[62] The remaining defect-rich region is often called the "virtual electrode" that facilitates the formation of conductive filaments in the subsequent set operation.

As mentioned earlier, in SNN applications, memristors are required to have high and low resistance states (neuron) and stable multi-resistance state (synapse) characteristics, that is, analog devices. Analog devices are easier to implement in the VCM mechanism, and the explanation for the multi-resistance state characteristics is mainly due to the different number or cross-sectional area of conductive filaments corresponding to different resistance states. In order to obtain stable and linear simulation characteristics, many studies have been carried out.[63–66] The 60 nm thick TaO_x thermal enhancement layer was believed to facilitate the formation of multiple conductive filaments, thereby achieving analog resistive properties.[67,68] Recently, 2048 conductance levels in a memristor were achieved through an electrical operation protocol to denoise the memristors for high-precision operations. The denoising process has been successfully applied to the 256×256 crossbars using the on-chip driving circuitry designed for regular reading and programming without extra hardware.[54]

Another VCM device has no conductive filament, which is based on interfacial redox reactions. Such a reaction involves oxygen vacancy exchange between two material layers. A typical example is the memristive device based on redox processes occurring at the electrode/oxide interfaces, leading to varying thicknesses of interfacial layers and, thus, different resistance states. Since such redox processes take place uniformly at the interface, it avoids the stochastic filament formation process and ensures good device-to-device and cycle-to-cycle uniformity. As shown in Figs. 7(d) and 7(e), for a $Pr_{0.7}Ca_{0.3}MnO_3$ (PCMO)/TiN-based device, in the high resistance state, a TiO_xN_y oxide layer is formed at the PCMO/TiN interface, which functions as an electron barrier that limits the current flowing through the device. When a positive voltage is applied to the anode, the reduction would first occur inside the TiO_xN_y layer, reducing its thickness, improving the device's conductivity to turn it into the low resistance state, and generating moveable oxygen ions in the layer. These oxygen ions would then drift from the TiO_xN_y layer to the PCMO layer under the driving force of the local electric field and oxidize the PCMO. If a negative voltage is applied to the anode, a reverse redox process would take place, which increases the thickness of the TiO_xN_y layer and turns the device from LRS to HRS.[55]

3.1.2. ECM

Another primary type of operation mechanism used for neuromorphic devices is ECM. In devices based on such a mechanism, the resistive change is realized by the migration of metal ions and the formation of a metallic conductive filament

between the electrodes. Metal ion migration-based devices typically possess a two-terminal metal/electrolyte/metal structure. Contrary to VCM-based devices, active metals are often adopted as the anode material or are doped into the electrolyte material in these types of devices to ensure metal ion migration under electric field control. In contrast, the cathode materials can still be inert metals, conductive nitrides, and oxides. The redox of metal atoms and ions at the electrodes and in the electrolytes is essential for the switching operation.[69,70]

In biological synapses, the accumulation and extrusion of Ca^{2+} in the pre-and postsynaptic compartments play a critical role in initiating plastic changes, as shown in Fig. 8(a). The ECM devices are designed to emulate this biological process in electronic devices in the use of diffusive metal cations (i.e., Ag^{+} and Cu^{2+}). [20]

In typical ECM devices, followed by the incorporation of cations, a positive voltage would be applied to the anode, which continuously oxides the anode surface, providing enough cations to form the conductive filament. At the same time, counter-reactions might happen simultaneously at the cathode/electrolyte interface to maintain charge neutrality. The cations accumulated at the cathode surface would be reduced to metal ions and grow larger. Under the control of the local electric field, the cations would migrate and gradually form the conductive filament. As shown in Fig. 8(c), the evolution of the conductive filament has been observed with different biases at different times. The same spontaneous rupture of the conductive filament and a consequent clustering process are also observed, which leads to current decay in the electrostatic performance of the ECM device.[71] Based on such ion dynamics, Wang *et al.* demonstrated a diffusive memristor, which could experience the same set process from high to low resistance. However, after the set voltage is removed, the conductance of such a device could spontaneously relax to a relatively high state, resulting in the realization of volatility, namely, short-term plasticity for neuromorphic computing. It has also been shown in Fig. 8(b) that prolonged or excessive stimulations with high-frequency pulses will eventually lead to an inflection of ECM device from facilitation to depression, an effect solely induced by an increased number of stimulation pulses at the same frequency, which is similar to bio-synapses.[20]

Generally, the operation of ECM devices also requires a forming process to incorporate the metal ions into the electrolyte layer, a set process to ensure metal ion migrations and the formation of the conductive filament, which leads to the switch from the HRS to the LRS, and a reset process to dissolute the filament and switch the device from the LRS to the HRS.

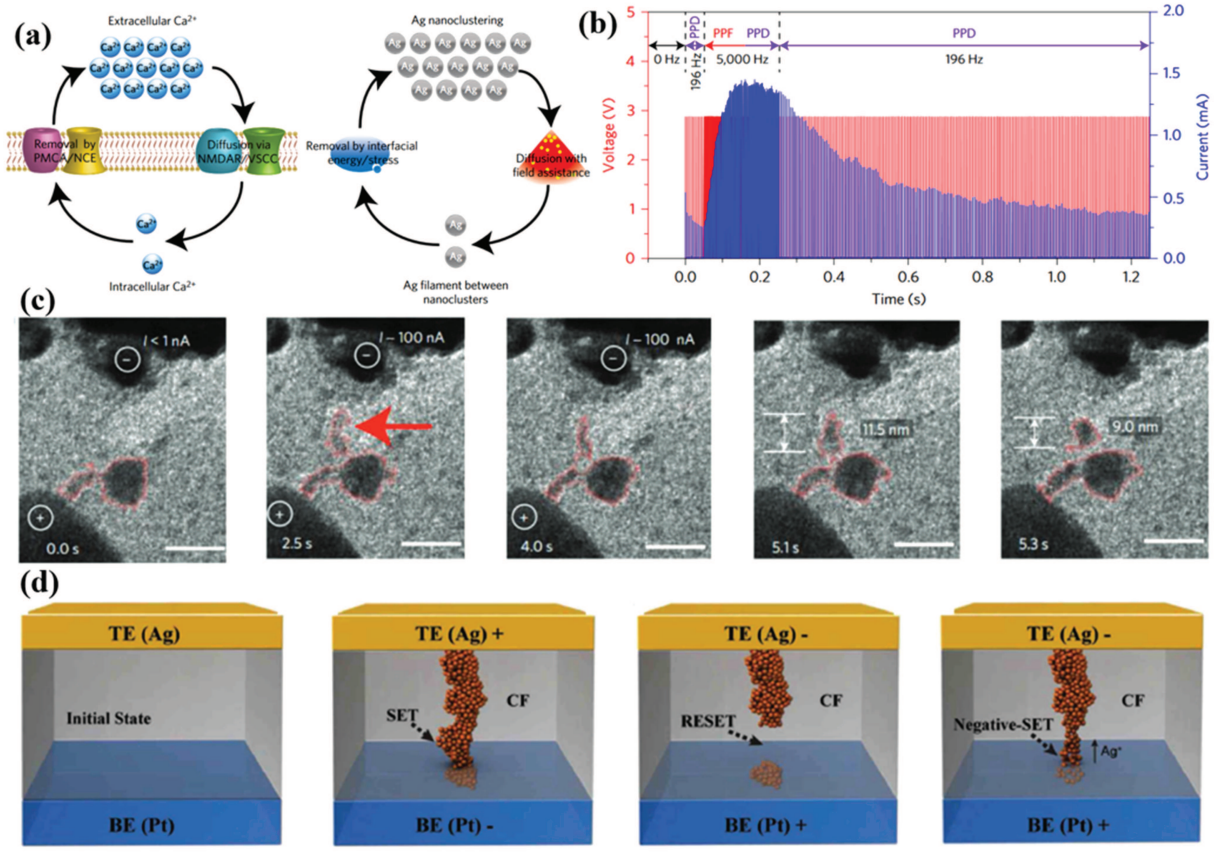


Fig. 8. ECM devices. (a) Schematic illustration of the analogy between biological Ca^{2+} and biomimetic Ag^{+} dynamics.[20] (b) Short-term synaptic plasticity of the ECM device.[20] (c) *In situ* HR-TEM observation of Ag diffusion process.[20] (d) Schematic illustrations of the switching mechanism of ECM device.[71]

In practice operation, negative-set processes may exist. As shown in Fig. 8(d), when a higher negative voltage is applied, the Ag precipitation in the bottom electrode as a local anode is oxidized and supplies Ag into the electrolyte.[71] Most of the Ag will move to the tip of the residual filament where the maximum electrical field exists in the electrolyte layer. When enough Ag is reduced to Ag atoms in the gap region, the filament will be reconnected to the bottom electrode, and the device will turn to LRS again. This contributes to the negative set process. Both the negative set and reset processes would increase the instability of the ECM cell and lead to computational failures.

In addition, the key benefits of memristor are their compactness with a $4F^2$ footprint and low operation voltage. These characteristics allow reductions in the hardware area and required electrical power of the neuromorphic system. Such memristor-based synapses and neurons can be monolithically integrable over pre-existing CMOS circuitry with the aid of a back-end-of-line approach. However, their use can lay a heavy burden on the back-end-of-line and prolong the turnaround time owing to the increased number of fabrication steps compared to the cointegration on a coplanar wafer surface.

3.2. PCRAM

PCRAM devices have a high resistance state and a low resistance state. The set and reset processes in PCRAM devices are based on the crystallization and amorphization of the phase change materials. As shown in Fig. 9(a),[72] by applying a low but wide voltage pulse, the set process can be realized. The growth of nanocrystals happens gradually until the amorphous region is finally transferred into a polycrystalline material. The PCRAM device gradually switches from the HRS to the LRS during the set process. During the reset process, a high but short voltage pulse is applied, melting part of the phase change materials. Such melted materials would cool down very quickly, forming an amorphous region. Such an amorphous region has lower electron density, which results in the resistive transition from the LRS to the HRS. From this working mechanism, stochastic spiking characteristics of synapse can be achieved. Although the PCRAM neuron has a low operation voltage and a high frequency, it requires additional circuits to detect the resistance and provide the appropriate reset pulse.

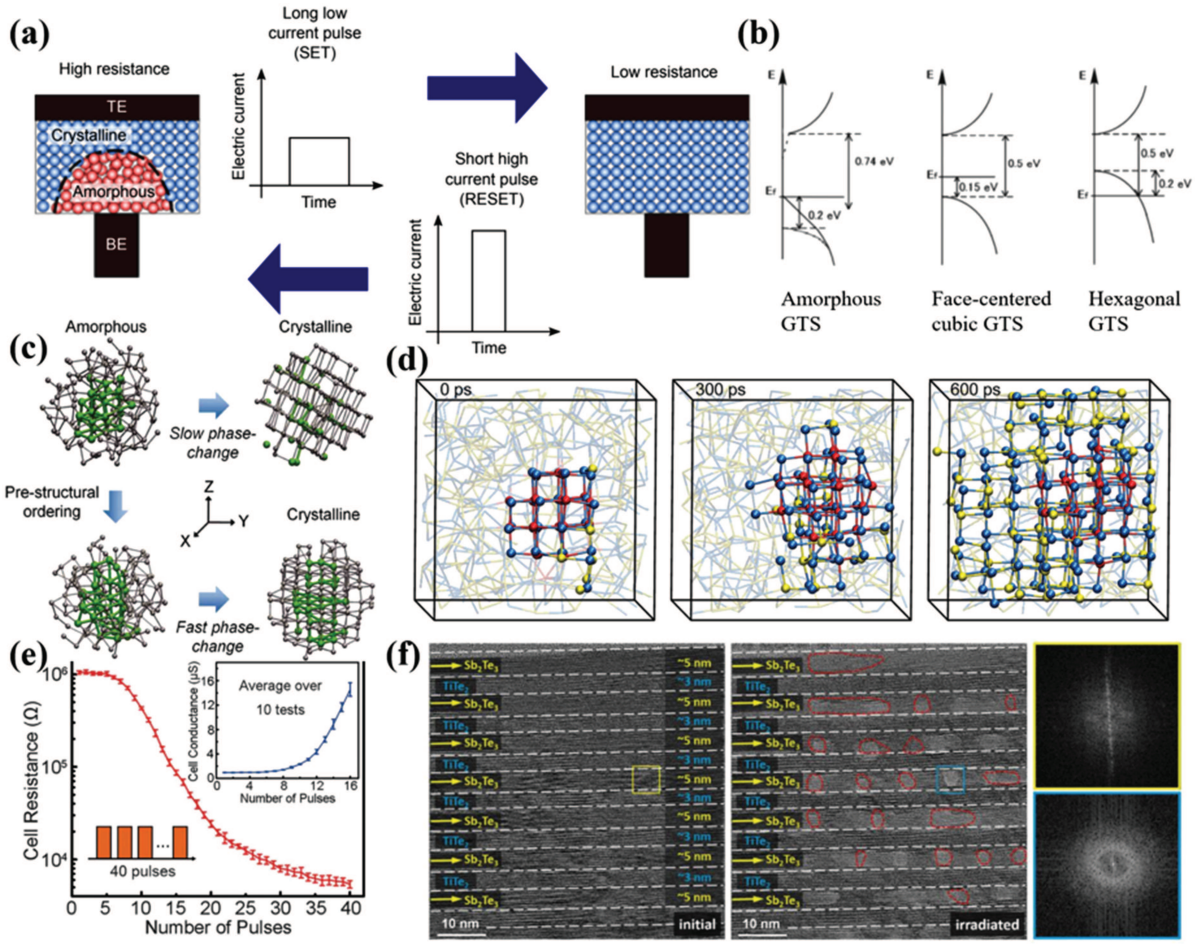


Fig. 9. PCRAM. (a) Schematic illustration of resistive switching in PCM device.[72] (b) Schematic band structure of the GST.[73] (c) Model configurations demonstrating the atomic rearrangements during the phase transition.[75] (d) Crystallization process of SST with a crystalline embryo in the middle at ~ 600 K.[76] (e) 40 pulse voltage of 0.8 V and 100 ns was operated for the cumulative SET.[77] (f) HR-TEM image for the phase-change heterostructure of $Sb_2Te_3/TiTe_2$ before and after irradiating with strong electron beams for 5 min.[77]

$Ge_2Sb_2Te_5$ (GST) is a typical phase change material whose electron conduction at both crystalline and amorphous states is shown in Fig. 9(b).[73] In both cases, the Fermi level lies close to the valence band, while the material conductivity is p-type. There is a valence band discontinuity between the crystalline and the amorphous GST. With the presence of structural vacancies in the crystal leading to acceptor-like traps, the pronounced band tail of the valence band and a high density of donor/acceptor defect pairs appear in the amorphous GST.[73,74] The microscopic origin of the crystallization in GST has been reported,[75] which illustrates that the fourfold rings of Ge inside the GST material accelerate the nucleation process and lead to the relatively high crystallization speed, as shown in Fig. 9(c). Proceeding nuclei inside the amorphous region of GST can bypass the incubation process and ensure sub-ns crystallization speed.

Besides GTS, other materials were explored for faster and more stable crystallization speed. The scandium antimony telluride ($Sc_{0.2}Sb_2Te_3$, SST) compound was designed to allow a writing speed of only 700 ps without preprogramming in a large conventional PCM device.[76] This ultrafast crystallization stems from the reduced stochasticity of nucleation through geometrically matched and robust scandium telluride (ScTe) chemical bonds that stabilize crystal precursors in the amorphous state. In DFMD simulations, the crystalline precursor made of ScTe cubes ($\sim$ 50 atoms in a four by four by four SST supercell made of 428 atoms) can stand robust against thermal fluctuations at $\sim$ 600 K in the absence of artificial constraint, serving as the center for subsequent crystallization (Fig. 9(d)).

New phase-change heterostructure that consists of alternately stacked phase-change and confinement nanolayers to suppress the noise and drift allows reliable iterative RESET and cumulative SET operations for high-performance neuro-inspired computing, as shown in Fig. 9(e).[77] To observe the phase change evolution in the W/Sb_2Te_3/TiN device, it was characterized by the HR-TEM under irradiation of intense electron beams. The HR-TEM results in Fig. 9(f) reveal the distinguishable regions in crystalline and amorphous states. Namely, the conductance of W/Sb_2Te_3/TiN phase change device was reconfigured by the irradiation of electron beam. The growth of the crystal in the amorphous region was dominated by the irradiation time. The resistance of the W/Sb_2Te_3/TiN memristor as a function of the number of the bias voltage pulse of 0.8 V and 100 ns showed a gradual SET behavior, in which the dynamic process was diagrammed as shown in Fig. 9(e). The iterative RESET and gradual SET for the W/Sb_2Te_3/TiN device demonstrate that the information storage is nonvolatile, making it a desirable emerging device for neuromorphic computing.

PCRAM appears as a promising candidate for neuromorphic computing due to its great maturity, excellent stability, low operation power, and promising scalability. Intel Xpoint Memory as an PCRAM with a double storage-selector stacked memory cell based on GTS materials demonstrated the technical feasibility of large-scale integration.

3.3. FeRAM

Ferroelectric tunnel junction (FTJ) comprises two electrodes separated by an ultrathin ferroelectric tunnel barrier, which is about 3–4 nm to allow electrons to tunnel through and maintain their ferroelectric nature. The reversal of spontaneous polarization in the ferroelectric tunnel barrier alters the internal electric field profile across the FTJ in conjunction with the different charge screening characteristics of the two electrodes composed of different metals, consequently changing the tunneling current. As shown in Fig. 10(a),[78] supposing the initial electronic potential barrier is rectangular, polarization charge effects would induce an asymmetric modulation of the electronic potential profile. The asymmetry of the electronic potential profile results in a higher average barrier height (Φ) when polarization points to the left (Φ_+) than when polarization points to the right (Φ_-). Since the tunnel transmission depends exponentially on the square root of the barrier height, the junction resistance will strongly depend on the polarization direction. The asymmetry between the two ferroelectric contact interfaces is thus essential to modulate the current transmission of the ferroelectric barrier.

Ferroelectric materials such as the $BiFeO_3$, (Ba, Sr)TiO_3 and (La, Sr)MnO_3 were extensively studied for the ferroelectric switching mechanisms and applications. As shown in Fig. 9(b),[76] experimental results demonstrated that the resistance change of the $BaTiO_3$/$La_{0.67}Sr_{0.33}MnO_3$ (LSMO) was dominated by the proportion of specific polarization direction in the tunneling layer.[81] As for $BiFeO_3$, combining scanning probe imaging in Fig. 10(c), electrical transport, and atomic-scale molecular dynamics, the conductance variations can be modeled by the nucleation-dominated reversal of domains. The 3D polarization configuration of the center-type domain structure in the $BiFeO_3$ nano-island was diagrammed in Fig. 10(d), in which the color arrows denoted the different polarization vectors.

The square $BiFeO_3$ nanoislands were fabricated further to study the polarization direction and DW switching behavior — topography of $BiFeO_3$ (001) thin film self-assembles as ordered nano-islands. The STEM images in Fig. 10(e) also reveal that the $BiFeO_3$ nano-island with a vertical 71° DW has head-to-head (H–H) polarization configuration, in which the CDW denotes the charged domain wall. The CDW with a $\sim$ 3–5 nm width is more significant than the neutral-charged DWs. Low- and high-conductance states of the $BiFeO_3$ nano-island have been obtained after applying a bias voltage of 1.5 V.

The switching behavior is very stable in 100 cycles between the two conductance states, and the resistance ratio of $> 10^3$ can be well maintained. Thus, the $BiFeO_3$ nano-island-based crossbar arrays with nonvolatile data storage can be used for synaptic devices.[79]

Ferroelectric neuron was also demonstrated with a leaky FeFET.[80] Compared to the nonvolatile devices for artificial synapse, the leaky FeFET utilizes accelerated degradation of polarization in a leaky ferroelectric hafnium zirconium oxide ($HfZrO_x$) layer, meaning that it can be used in neuron devices with leaky characteristics. The voltage pulse applied to the gate of a leaky FeFET can change the polarization of the ferroelectric layer, and the drain current as the form of I_{out} flows upon the application of a certain number of pulses. When the voltage pulse is not applied in the resting state, I_{out} is reduced naturally owing to the rapid degradation of polarization in the leaky ferroelectric layer. Therefore, it can act as an LIF neuron in SNN.

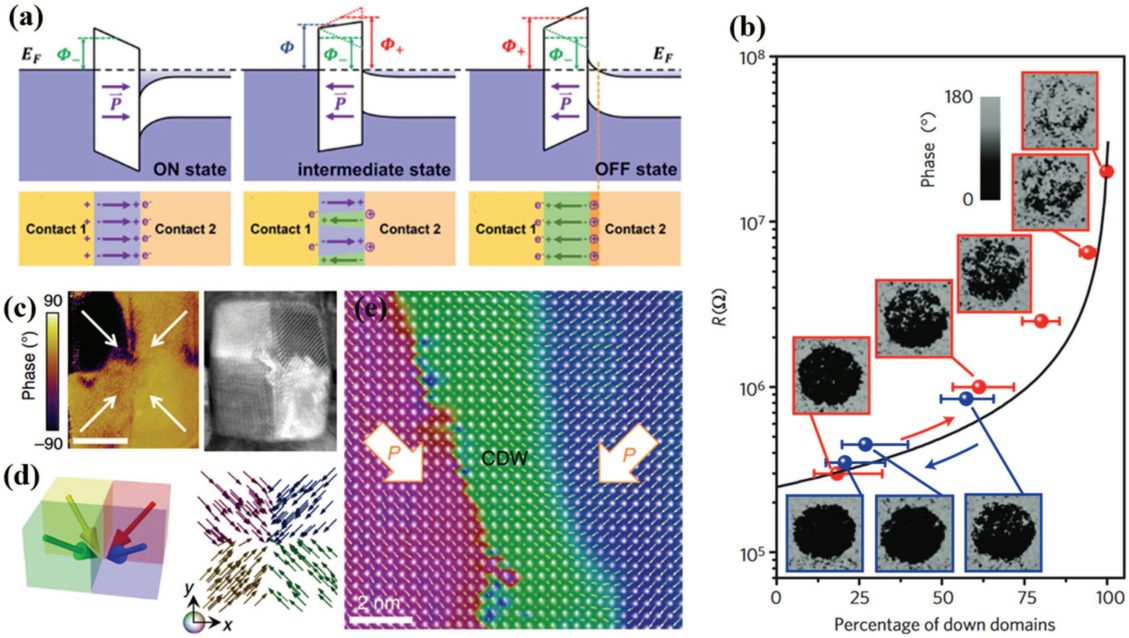


Fig. 10. FeRAM. (a) Band diagrams and schematics of FTJ in ON state, intermediated state, and OFF state.[78] (b) Resistance versus the relative fraction of downward polarization domains for the $BaTiO_3/La_{0.67}Sr_{0.33}MnO_3$ (LSMO) junction.[76] (c) In-plane PFM image (left) and annular dark-field image showing a square-shaped $BaFeO_3$ (BFO) nano-island containing four sections (right).[79] (d) Schematic of the 3D polarization configuration of the center-type domain structure in the BFO nano-island.[79] (e) HAADF STEM image of the H–H DW region.[79]

The $HfZrO_x$ is a commonly used material for ferroelectric devices and has extraordinary compatibility with the conventional CMOS technique because the HfO_2 dielectric has been widely used as a high-k gate dielectric in the current CMOS technology. Hence, it is believed to be a promising candidate for future large-scale SNN.

3.4. MRAM

Devices based on ferromagnetic materials have been systematically studied, providing excellent potential for artificial neuromorphic applications. As synaptic devices, MRAMs have relatively good performance in the read and write operations with the high speed ($\sim$ ns), low energy consumption ($\sim$ pJ) and excellent endurance ($> 10^8$) characteristics. When the input pulses are applied, spike trains can be generated with a bias-dependent duty cycle, allowing it to behave as a stochastic neuron for an SNN. Among MRAMs, spin transfer torque (STT) magnetic tunneling junction (MTJ) and spin–orbit torque (SOT) devices have attracted significant attention in the past few decades as promising candidates for neuromorphic computing.

3.4.1. STT

A unit MTJ has two metallic ferromagnetic layers separated by a tunnel barrier, as shown in Fig. 11(a).[78] The tunnel barrier is a paraelectric insulating layer. The conductance of the MTJ is dominated by the magnetization directions of the two metallic ferromagnetic layers. The thicker ferromagnetic layer is used as the fixed layer, also termed the pinned layer (PL) where the spin polarization is pinned to a particular direction. The other thin layer is the free layer (FL), where the magnetization direction is changed by the injection of current. Importantly, intermediate resistance states between parallel and antiparallel magnetic directions are obtained for the free layer with injection current. The parallel magnetization process offers a higher probability of electron tunneling through the barrier of the MTJ than the antiparallel configuration. Thus, two distinct resistance states, the HRS of the antiparallel and LRS of the parallel structures, are observed in STT devices.[84]

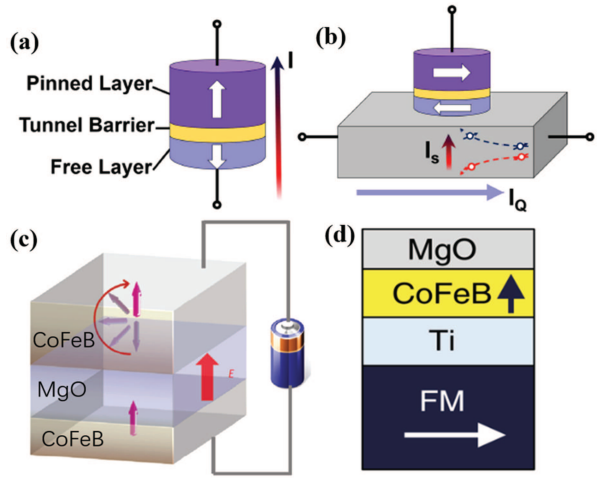


Fig. 11. MRAM. (a) Schematics of MTJ consists of two ferromagnets, the free layer and the pinned layer separated by a tunneling oxide barrier.[78] (b) Spin–orbit torque is generated on a nanomagnet due to charge current flow (I_Q) through an underlying heavy metal layer due to the spin-Hall effect.[78] (c) Electric-field-assisted switching in a CoFeB/MgO/CoFeB MTJ.[82] (d) External-magnetic-field-free switching of perpendicular magnetization in a CoFeB/Ti/CoFeB/MgO SOT device.[83]

The ferromagnetic fixed layer is spin-polarized in the defined magnetic direction $\hat{m}$. The electron injected into the PL externally has the same spin direction due to the magnetization of $\hat{m}$. If the concentration of injected electrons is low and not able to switch the magnetization direction of the FL, the MTJ remains in the HRS. If the concentration is high, a large number of electrons with the same magnetization of $\hat{m}$ will tunnel through the oxide insulator layer. The torque is induced by the difference between the magnetization of $\hat{m}$ and FL. In this case, the magnetization of the FL gradually orients in the direction of $\hat{m}$. The intermediate conductance states of MTJ are useful for neural networks and analog computing. When the magnetization of the FL has the same direction of $\hat{m}$, namely, the magnetization parallel, the injected electron freely tunnels through the MTJ, the device switches from the HRS to the LRS.[85]

The tunnel barrier is an insulator with a thickness of about 1 nm made of a crystalline MgO material, which is the most-suited tunnel barrier material up to date.[86,87] CoFeB alloys are the state-of-the-art ferromagnets employed in MTJs.[88] In these devices, the thickness of the CoFeB free layer has to be thinner than about 1.5 nm to favor the out-off plane orientation of its magnetization.

Although the read power consumption for MTJ could be extremely low, which is almost zero, the power consumption for writing could be much higher, making the power consumption up to $\sim$ 100 fJ/spike, which shows no significant advantage to other neuromorphic devices. Hence, many efforts were made to resolve the relatively large write power.

The electric field can modify the exchange coupling between two ferromagnetic layers through a non-magnetic spacer, which enables the magnetization switching without an external magnetic field. As shown in Fig. 11(c), electric-field-assisted reversible switching in CoFeB/MgO/CoFeB MTJ can be manipulated by voltage pulses associated with much smaller current densities.[82]

3.4.2. SOT

SOT-induced magnetization switching shows promise for realizing ultrafast and reliable spintronics devices. Typical structures of the SOT device employ a nonmagnet/ferromagnet (NM/FM) bilayer. The spin Hall effect (SHE) and Rashba effects originate from spin–orbit coupling within the NM layer and at the FM/NM interfaces, respectively, responsible for the switching. Insulator–ferromagnet–heavy metal (I-FM-HM) multilayer structures in Fig. 11(b) own much greater spin injection efficiencies due to strong spin–orbit interaction. When a charge current flows through the underlying HM, SOT is generated at the FM–HM interface.[78] Although the cause of SOT can be attributed to two possible origins, SHE is usually the dominant underlying physical phenomenon. Due to the flow of charge current through the HM, electrons with opposite spins scatter on the top and bottom surfaces of the HM. The spin-polarization is orthogonal to both the directions of the charge current and injected spin current. These electrons repeatedly experience spin-scattering while traveling through the HM, transferring multiple units of spin angular momentum to the FM lying on top. The magnitude of injected spin current density (J_s) is proportional to the magnitude of input charge current density (J_q), with the proportionality factor being defined as the spin-Hall angle ($\theta_{SH} < 1$). Hence, the input charge to spin current conversion is governed by the following relation:

$$I_s = \theta_{SH}\left(\frac{W_{FM}}{t_{HM}}\right) I_q, \tag{23}$$

where I_s and I_q are the input spin current and charge current, respectively, W_{FM} is the width of the FM lying on top of the HM, and t_{HM} is the HM thickness. By ensuring $W_{FM} > t_{HM}$, high spin injection efficiencies greater than 100% ($I_s > I_q$) can be achieved.

This has been suggested to induce switching of the perpendicular magnetic component within the trilayers of in-plane CoFeB/Ti/perpendicular CoFeB (Fig. 11(d)).[83] Meanwhile, when additional crystal or magnetic symmetry breaking is introduced, perpendicular and longitudinal spins can also be generated, e.g., in low symmetry crystals, non-collinear antiferromagnetic crystals with magnetic asymmetry or magnetic SHE, and also some collinear antiferromagnet crystals with spin conversions or spin splitting effect.

In addition, considering that the MRAM should be formed in the back-end-of-line, it can also give rise to fabrication complexity, especially on the back-end-of-line, and delay the turnaround time owing to the increased number of the fabrication steps.

4. In-sensor computing

With the rapid development of information technology, the volume of data has shown a blowout growth, which also puts forward higher requirements on the computing power, and the demand for the application of artificial intelligence in various fields is increasing, especially when dealing with complex, perception- and decision-related tasks, which require more complex and efficient computational models. However, with the unfortunate fact of the gradual failure of Moore's law, an efficient process structure must be proposed. Biological neural systems, with their powerful parallel computing and adaptive learning capabilities, are able to process a large number of complex and changing signals in the surrounding environment quickly and in real time with low energy consumption and high arithmetic power, which makes neuromorphic computing a possible solution for the development of AI.

Inspired by biological neural systems, in-sensor computing architecture has been proposed and artificial neurons have been prepared triumphantly. Researchers have successfully simulated the biological senses to realize the processing of complex signals, including touch, smell, hearing, and vision, which acquire the most information based on the related research results.

In 2017, Shulaker *et al.* from Stanford University prepared a 3D vertically stacked chip that interconnects sensors and processing circuits in the vertical direction, realizing the integrated processing of sensing, storage, and arithmetic in a single chip, which shortens the data handling distance through vertical interconnections and enables *in-situ* processing of data, and ultimately succeeds in distinguishing nitrogen from six types of vapors: lemon juice, white vinegar, rubbing alcohol, vodka, wine and beer.[89] In 2012, Hsieh *et al.* from National Tsing Hua University, Taiwan of China developed a low-energy SNN chip with odor recognition function, which used circuits to simulate the functions of relevant cells and structures in the olfactory system, such as mitral valve cells, STDP synapses, sebaceous cells, etc. The chip was fabricated by TSMC's 0.18-μm CMOS process, with a chip size of 1.78 mm^2. The average power consumption was 3.6 μW, and the test accuracy was 87.59%.[90]

In 2017, TermehYousefi *et al.* at Kyushu Institute of Technology, Japan, fabricated a hemispherical artificial fingertip based on nanocomposites to enhance the tactile sensing system of a humanoid robot. Ripe and unripe tomatoes were categorized by recording their metabolic growth as a function of resistivity change under controlled indentation force.[91] In 2018, Yeongin *et al.* from Stanford University successfully simulated tactile perception using flexible electronics, with a ring oscillator that converts pressure information (1–80 kPa) into action potentials (0–100 Hz), and a layered structure of the device that detects the movement of an object, as well as combining the analysis of pressure signals to differentiate Braille characters.[92]

In 2018, Wang *et al.* reported a nm-scale device that enables synaptic plasticity and proposed a brain-like spatiotemporal algorithm that exploits the sensitivity of SNN networks to the arrival time of impulse spikes to successfully simulate the azimuthal angle detection of sound, enabling the simulation of biological hearing.[93] In 2018, Wang *et al.* from Zhengzhou University proposed a novel developmental network (DN) that simulates the human auditory system and constructed an artificial auditory model for speech recognition. The highest recognition rate for English word speech reached 93.60%.[94]

In 2021, Tan *et al.* from Aalto University, Finland introduced a bio-inspired spiking multisensory neural network, which integrates artificial vision, touch, hearing, simulated olfaction, and gustation with cross-modal learning via artificial neural networks. In one of the artificial olfactory and gustatory systems, olfaction and taste are simulated by nine (etheric, aromatic, sweet, spicy, oily, burnt, sulfuric, rancid, metallic) and five (sweet, sour, salty, bitter, fresh) receptor potentials, respectively. Spike encoders convert the analog voltages into light spikes, and photosensitive transistors decode and memorize the sensory information.[95]

4.1. Human vision system structure

Eighty percents of the information that the human body acquires comes from the visual system, so the importance of vision cannot be overstated. The biological visual system consists of the retina, optic nerve, and brain. The rods and cone neurons in the retina capture the visual information and transform light signals into electronic signals. Subsequently, the optical nerve can transfer electronic signals to the brain, where it acts as the information processor, achieving classification, recognition, learning, and memory. Through selective attention, the brain performs efficient information processing.

As briefly mentioned above, the retina plays a vital role in visual. When the light arrives, photoreceptors in the retina are activated, creating an electronic signal which is conveyed to the optic nerve and then to the brain, as shown in Fig. 12. These important functions are carried out by specific cells in the retina. Cone cells can sense light and distinguish color. According to the sensitive wavelength, they can be broadly divided into three categories: red, green, and blue. Rod cells are sensitive to low light. Even a single photo can cause cell excitation. Horizontal cells play a connecting role, adjust the brightness of the signal, and enhance the outline of the scene at the edge of the visual image. Bipolar cells convert sustained graded potentials into transient neural activity. Nodal cells are responsible for the encoding and transmitting electrical impulses of visual information.

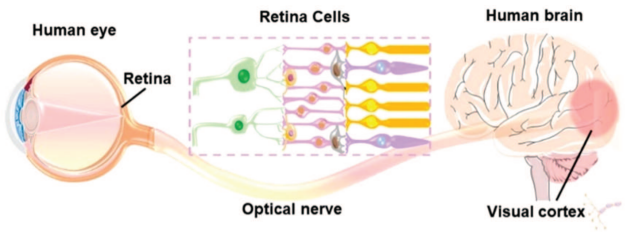


Fig. 12. Human vision system.[96]

Among the cerebral cortex, the visual cortex is primarily responsible for processing visual information. There are five critical areas in the visual cortex. V1 can process visual features such as direction motion, spatial frequency, parallax, color, etc. V2 can slightly complex modulation of positioning, spatial frequency, color, etc. And V2 could moderate the modulation of complex shapes. V3 is the second most important motion processing area responsible for the motion processing of the entire side field of view. V4 could handle simple geometric shapes but not complex shapes like faces. V5 is responsible for processing complex visual motion stimuli, observing the complex motion of individual points in the visual field caused by the relative movement to the environment.

4.2. Sensor pre-processing structure

A portion of the information is pre-processed on the retina and then transmitted to the brain via the optic nerve, as shown in Fig. 13. In this way, the human eye accomplishes more complex tasks such as recognition and classification in a more energy-efficient and effective manner. Inspired by the way the human eye processes, researchers have also front-loaded a portion of the data preprocessing to the sensor unit, and the data obtained by the sensor is passed through the bus to the back-end processor to realize advanced functions, such as the recognition and classification of images and audio, and other parts.

In 2018, Li *et al.* designed a one transistor and one memristor (1T1R) structure, achieving analog signal and image processing, forming an array of 128×64.[97] In 2019, based on CMOS image sensors through 2D on-chip integration, Chu *et al.* assembled sensing, storage, and arithmetic functions onto the same PCB board, which is able to simulate LIF neurons and realizes simple image recognition functions through SNNs.[98] In 2022, Valenzuela proposed a smart image sensor using on-pixel memory based on a 0.35 μm CMOS process, achieving object positioning.[99]

Summarize the above architecture, through the logic circuit for each detector unit data processing scheme, there is a common problem: larger pixels and a smaller fill factor, which seriously limits the resolution of the imaging chip. What is worse, the more complex the functions to be accomplished, the more components the logic processing circuits contain, and the resolution is even more limited. To solve it and further reduce the power and time loss caused by the transmission of large amounts of data, a three-dimensional integrated vision chip architecture is proposed as an alternative to the two-dimensional planar architecture.

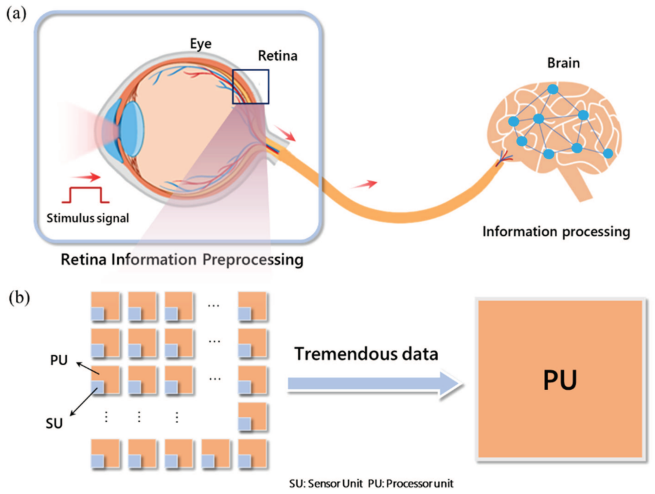


Fig. 13. Retina-inspired sensor preprocessing structures.

In 2018, Nose *et al.* developed a high-speed vision chip using 3D stacking technology, and combined it with column parallel processing architecture, finally realizing high sensitivity, high resolution, and high accuracy target localization.[100] In 2017, Shulaker *et al.* used three-dimensional integration to realize an integrated sense–storage–accounting chip consisting of more than one million resistive random-access memory cells and more than two million carbon nanotube field-effect transistors, which is capable of realizing large-scale parallel, *in-situ* processing of data.[101] In 2022, Hirata *et al.* designed a back-illuminated stacked sensor to achieve a frame rate of 1000 fps and a resolution of $4k \times 4k$.[102]

Although the form of 3D stacking reduces the data transmission distance and improves the transmission bandwidth to a large extent, there is still the problem of transport limitation of bandwidth, i.e., the "transport wall" due to the lack of essential innovations in the architecture, and the solution of 3D stacking also brings troubles to the manufacturing process.

4.3. Integrated sensing–memory–processing structure

Based on the analysis above, the conclusion is that a new architecture that is closer to the human's eye's operation mode needs to be proposed, thus further increasing speed and reducing energy consumption. It is worth noting that the human eye processes pulsed signals, and every cell in the retina is capable of data acquisition and processing. In contrast, data transportation in the retina is highly parallelized. This highly parallel, *in-situ* processing, sensing the amount of change working model, has two significant advantages: high speed and low energy consumption. A human eye-inspired imaging chip structure is shown in Fig. 14.[103]

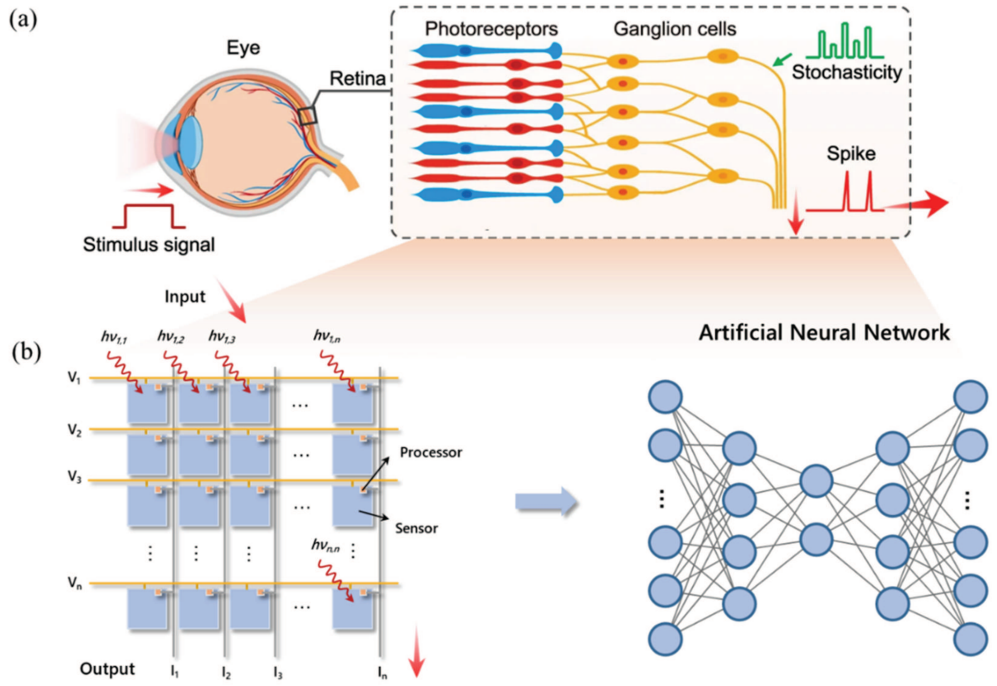


Fig. 14. Retina-inspired integrated sensing–memory–processing structure.[103]

In 2021, inspired by biological eyes, Meng *et al.* fabricated an artificial retina based on 2D Janus MoSSe, which could simulate visual perception functions with electronic/ion and optical comodulating. In the formation of a faradic electric double layer (EDL) at the metal-oxide/electrolyte interfaces, it could realize synaptic weight changes. The 2D MoSSe-based electronic device realized the integration of visible information sensing–memory–processing.[96] In 2022, Yu *et al.* proposed a programmable ferroelectric bionic vision hardware with selective attention for high-precision image classification. A programmable ferroelectric bionic vision hardware with selective attention was fabricated using quantum dots (QDs) and ferroelectric material.[104] In 2022, Lee *et al.* demonstrated a heterogeneously integrated 1-photodiode and 1-memristor (1P-1R) crossbar for in-sensor visual cognitive processing, the trained weight values are applied as an input voltage to the image-saved crossbar array, realizing the in-sensor computing paradigm.[105] In 2018, Seo *et al.* demonstrated an optic-neural synaptic device by implementing synaptic and optical-sensing functions together on h-BN/WSe_2 heterostructure. This device mimics the colored and color-mixed pattern recognition capabilities of the human vision system when arranged in an optic-neural network.[106] In 2023, Wang *et al.* presented a retina-inspired mid-infrared (MIR) optoelectronic device based on a two-dimensional (2D) heterostructure for simultaneous data perception and encoding.[103] In 2017, Wang *et al.* used the dispersion and aggregation behavior of silver atoms to mimic the influx and extrusion of Ca^{2+}, thus simulating biological synapses behaviors.[20] In 2023, Tan *et al.* utilized a photoresistor array to do processing of input pulse signals to achieve dynamic vision. This dynamic processing capability in the sensor eliminates redundant data streams and facilitates real-time perception of moving objects by dynamic machine vision.[107] A general trend is that researchers are now focusing on the development of photonic synaptic devices in an attempt to advance retinal simulation through the simulation of synaptic properties. Some metal oxides, sulfide heterojunctions, two-dimensional materials, etc. are widely used in the realization of photonic synaptic devices due to their great potentials. They successfully model the short-term plasticity and long-term plasticity of synapses to achieve effective modulation of optical signals and real-time image processing, but there are still some drawbacks, the perception of wavelengths that are not in the visible light range and high energy consumption due to large light intensity.[108,109,111–116] A more specific performance comparison of the AI vision chips mentioned above is shown in Table 1.

According to the comparison of the above table, it can be found that the current AI vision chips generally have the problem of low sensitivity, which requires substantial incident optical power to respond, so further improving the sensitivity is particularly important for improving the AI vision chip performance. This requires a higher sensitivity of the sensor. Among the many competitors, the single-photon avalanche diode (SPAD) stands out due to its excellent weak light detection capabilities. The SPAD can detect optical signals at the single photon level and can also process pulsed signals. These features are well suited for neuromorphic computing and important for developing high-performance AI vision chips.

Table 1. AI vision chip performance comparison.

Date	Key technology	Wavelength	Input light intensity
2012[114]	multilayer MoS_2	455–850 nm	50 mW/cm^2
2015[113]	monolayer MoS_2	532 nm	20 mW/cm^2
2017[111]	ITO/ZnO/PDR1A/Al heterojunction	514 nm	180 mW/cm^2
2018[112]	ITO/Si-NC/Al heterojunction	375 nm	18 mW/cm^2
2018[106]	h-BN/WSe_2	625 nm, 532 nm, 406 nm	18 mW/cm^2
2018[109]	ferroelectric/electrochemical modulation	550–850 nm	10.80 mW/cm^2
2019[108]	B-Si NCs	1342 nm	30 mW/cm^2
2019[110]	ITO/Nb:$SrTiO_3$ heterojunction	459 nm	30 mW/cm^2
2019[115]	ITO/MoO_x/Pd	365 nm	150 W/cm^2
2022[105]	integrated 1P-1R	532 nm	67 mW/cm^2
2022[116]	TiAu/GaO_x/SiO_2	254 nm	450 W/cm^2
2023[107]	2D $MoTe_2$	MIR	130 mW/cm^2

1P-1R: 1-photodiode and 1-memristor.

4.4. Single-photon avalanche diode device

An avalanche detector is good at detecting weak light signals because of its internal gain mechanism. Especially those working in the Geiger mode can achieve single photon level detection with an internal gain up to 10^6. This amazing result gives us much confidence to achieve the target of single-photo imaging. SPAD array could achieve a high degree of parallel operation and solve the problem of slow CCD reading speed. So if high-quality large-scale SPAD arrays could be made successfully one day, artificial intelligence imaging systems based on them could make a significant breakthrough in faint light imaging, which is of great significance for breaking through the limits of the human eye.

SPAD operates above the breakdown voltage when the depletion area forms a strong electric field (electric field strength up to kV/cm magnitude). When a photo hits the detector, the photoelectric effect produces an electron–hole pair. The electron (hole) rushes under the action of a high electric field, gains enough energy, and collides with the lattice to produce a new electron (hole). This chain reaction occurs continuously and is known as the collisional dissociation effect, where the number of carriers produced grows exponentially, culminating in a macroscopically detectable current. Above is the whole process of multiplication. However, the avalanche multiplication process is autonomous and cannot be quenched autonomously. Continued high current through the PN junction depletion region will lead to detector damage. And the detector in the self-sustaining state cannot sense new photon incidence, so the necessary quenching mechanism needs to be introduced to terminate the multiplication process and return the sensor to the initial state. The SPAD's avalanche-quenched mode of operation causes it to output a pulsed signal (independent of whether the input signal is pulsed or not). The time for the quenching and recovery process is not negligible, which poses a challenge for high-speed imaging. Efficient quenching mechanisms need to be proposed to improve the detector response speed.

Another effect that can trigger an avalanche is known as after-pulsing. When an avalanche occurs, the PN junction is flooded with charge carriers, and trap levels between the valence and conduction band become occupied to the degree that is much greater than that expected in a thermal-equilibrium distribution of charge carriers. After the SPAD has been quenched, there is some probability that a charge carrier in a trap level receives enough energy to free it from the trap and promote it to the conduction band, which triggers a new avalanche. Thus, depending on the quality of the process and the exact layers and implants that were used to fabricate the SPAD, a significant number of extra pulses can be developed from a single originating thermal or photo-generation event. The degree of after-pulsing can be quantified by measuring the autocorrelation of arrival times between avalanches when a dark count measurement is set up. Thermal generation produces Poissonian statistics with an impulse function autocorrelation, and after-pulsing has non-Poissonian statistics.

With after-pulsing, a SPAD cannot effectively detect single photons continuously, which directly leads to reduced sensitivity and detection speed. So, we must suppress the after-pulsing by improving the material quality, inserting an effective quench mechanism, and so on.

The leading edge of a SPAD's avalanche breakdown is particularly useful for timing the arrival of photons. This method is useful for 3D imaging, LIDAR, and is used heavily in physical measurements relying on time-correlated single photon counting (TCSPC). However, dedicated circuits such as time-to-digital converters (TDCs) and time-to-analog (TAC) circuits are required to enable such functionality. Two general processes complicate the measurement of a photon's arrival. The first is the statistical fluctuation in the photon's arrival time, which is a fundamental property of light. The second is the statistical variation in the detection mechanism within the SPAD due to a) depth of photon absorption, b) diffusion time to the active p–n junction, c) the buildup statistics of the avalanche, and d) the jitter of the detection and timing circuitry.

The amplitude of the avalanche pulse also impacts the imaging quality, as the intensity of the detector response corresponds to the power of the imaging at that point. Only if the amplitude of the avalanche pulse is consistent can the detector response truly reflect the image's intensity distribution. Nevertheless, the reality is that even with single photon incidence (well-modulated), we still observe fluctuations in the pulse amplitude. The fluctuation of the pulse amplitude may result in poor imaging quality because the random nature of the avalanche causes the detector to not be fully charged during the recovery process. We also need to find ways to suppress pulse fluctuations.

4.5. Future prospect of vision chips

In contrast to human vision systems, vision chips usually use 2D imagers, which would lose the depth information.[117] But depth information plays a vital role in many visional processing tasks, so 3D high-speed imaging is essential to promote the development of artificial intelligence imaging chips, which can be applied for fields such as high-speed machine vision, high-speed object recognition, high-speed target tracing, high-speed machine interaction, etc. in the future. To achieve 3D imaging, 3D vision sensors based on the time-of-light (TOF) principle are flourishing to meet this requirement.[118]

The traditional sense–computation separation structure has a large amount of data redundancy. Data transmission and redundant data processing consume a lot of time, limiting the high frame rates of imaging devices. Meanwhile, this processing is computationally intensive, demanding high arithmetic power, power consumption, and hardware complexity.[119] Inspired by the way living organisms process information, the development of low-energy, event-driven devices aims to achieve higher levels of artificial intelligence. Since SPADs are event-driven and asynchronous by nature, it is determinant that they will play an essential role in developing artificial intelligence imaging chips. SPAD imagers with high sensitivity, high frame rate, and large integration are the current development direction, and the realization of this goal will undoubtedly elevate the artificial intelligence imaging technology to a higher level.

5. Summary

Overall, neuromorphic computing aims to achieve low-power and high-speed computing by simulating the physical architecture and information processing methods of the human brain, which is expected to break through the von-Neumann bottleneck of traditional computing structure. When the human brain processes information, neurotransmitters from the pre-synaptic act on the post-synapse in the form of pulses, and neurons store and process information *in situ* through synaptic plasticity. Based on these two characteristics, SNN and in-sensor computing have been proposed to mimic the human brain. The former is based on signal transmission mode, while the latter is based on hardware architecture. Both of them make the present computing structure more brain-like.

As the third-generation artificial neural network, SNN exhibits the innate advantage of low-energy consumption and high-energy efficiency. Therefore, SNN can play to its strength in energy-sensitive applications such as mobile devices, the internet of things (IoT), etc. Additionally, given SNN's excellent performance in time-series information, it could excel in processing tasks related to time, like motion detection and biomedical signal processing. The implementation of SNN hardware could better solve several cognitive problems as it mimics the brain's processing, which might lead to more efficient AI systems. But the training methods of SNN are limited and there is a long way to go on extensive deep networks. It is believed that the architecture and algorithms of SNN will be constantly improving with the in-depth research of brain science. Memory devices based on new mechanisms and materials make the hardware implementation of large-scale SNN possible. In the future, bionics will give us more inspiration to process information, let us reveal the nature of event-driven mode, and give the SNN a clearer physical background.

In-sensor computing is a bio-inspired, emerging computing paradigm that reduces or eliminates the need to transfer data from the sensor to a central processing unit by performing data processing near or within the sensor. This technique can dramatically reduce the amount of data transfer, reduce computational complexity, increase the speed of data processing, and help save energy and improve safety. Based on the above advantages, in-sensor computing technology can play a big role in scenarios that require real-time processing of big data.

With the development of the IoT, the number of sensor nodes has proliferated, generating massive amounts of data. In-sensor computing allows real-time processing of data at the sensor nodes, reducing the challenges posed by data transmission, and improving the responsiveness and efficiency of the system. In the blueprint of smart city, a large number of sensors are needed to collect environmental data, such as air quality and traffic flow. The use of in-sensor computing technology can ensure the speed of data processing while reducing the resources and time required for data transmission, and improve the efficiency of intelligent management of the city. Factory automation and smart manufacturing are core components of industry 4.0. In these systems, sensors for monitoring various parameters on the production line are essential. Calculations in the vicinity of the sensors allow for real-time responses and adjustments to the production process, improving productivity and quality. In the field of healthcare, the initial processing of

data on sensors can provide real-time feedback on the health of the human body and transmit only the necessary data to the central processing system, improving efficiency while protecting personal privacy. Self-driving cars require real-time processing of data from a variety of sensors, such as radar, cameras, and LIDAR. In-sensor computing increases the speed of data processing and reduces latency, which enhances the responsiveness of the self-driving system. The security field usually needs to quickly process a large amount of image and video data to complete the task of face recognition, hazards recognition, etc. In-sensor computing can greatly relieve the pressure of data transmission through data compression and real-time processing, and is expected to realize the real-time processing of hundreds of millions of pixels to realize real-time monitoring.

To sum up, the ever-advancing neuromorphic computing will make information processing more intelligent and efficient, ushering in the rapid development of AI in the post-Moore era.

Acknowledgements

Project supported in part by the National Key Research and Development Program of China (Grant No. 2021YFA0716400), the National Natural Science Foundation of China (Grant Nos. 62225405, 62150027, 61974080, 61991443, 61975093, 61927811, 61875104, 62175126, and 62235011), the Ministry of Science and Technology of China (Grant Nos. 2021ZD0109900 and 2021ZD0109903), the Collaborative Innovation Center of Solid-State Lighting and Energy-Saving Electronics and Tsinghua University Initiative Scientific Research Program.

References

[1] Zhang C and Lu Y 2021 *Journal of Industrial Information Integration* **23** 100224
[2] Marković D, Mizrahi A, Querlioz D and Grollier J 2020 *Nat. Rev. Phys.* **2** 499
[3] Maass W 1997 *Neural Networks* **10** 1659
[4] Tavanaei A, Ghodrati M, Kheradpisheh S R, Masquelier T and Maida A 2019 *Neural Networks* **111** 47
[5] Wulfram Gerstner, Richard Kempter, J. Leo van Hemmen, Hermann Wagner 1996 *Nature* **383** 76
[6] Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg A C and Fei-Fei L 2015 *Int. J. Comput. Vis.* **115** 211
[7] Krizhevsky A, Sutskever I and Hinton G E 2017 *Commun. ACM* **60** 84
[8] Hinton G, Deng L, Yu D, Dahl G E, Mohamed A, Jaitly N, Senior A, Vanhoucke V, Nguyen P, Sainath T N and Kingsbury B 2012 *IEEE Signal Processing Magazine* **29** 82
[9] Girshick R, Donahue J, Darrell T and Malik J 2014 2014 *IEEE Conference on Computer Vision and Pattern Recognition*, June 24–27, 2014 Columbus, USA, p. 580
[10] Long J, Shelhamer E and Darrell T 2015 2015 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 8–10, 2015, Boston, USA, p. 3431
[11] Girshick R 2015 2015 *IEEE International Conference on Computer Vision*, December 13–16, 2015, Santiago, Chile, p. 1440
[12] Ren S, He K, Girshick R and Sun J 2017 *IEEE Transactions on Pattern Analysis and Machine Intelligence* **39** 1137
[13] He K, Gkioxari G, Dollár P and Girshick R 2017 2017 *IEEE International Conference on Computer Vision*, October 22–29, 2017, Venice, Italy, p. 2980
[14] He K, Zhang X, Ren S and Sun J 2016 2016 *IEEE Conference on Computer Vision and Pattern Recognition*, June 26–July 1, 2016, Las Vegas, US, p. 770
[15] Szegedy C, Vanhoucke V, Ioffe S, Shlens J and Wojna Z 2016 2016 *IEEE Conference on Computer Vision and Pattern Recognition*, June 26–July 1, 2016, Las Vegas, USA, p. 2818
[16] Geirhos R, Michaelis C, Wichmann F A, Rubisch P, Bethge M and Brendel W 2019 arXiv: 1811.12231 [cs.CV]
[17] Manassi M, Liberman A, Chaney W and Whitney D 2017 *Sci. Rep.* **7** 1971
[18] Fitzpatrick D 2000 *Current Opinion in Neurobiology* **10** 438
[19] Pi S, Li C, Jiang H, Xia W, Xin H, Yang J J and Xia Q 2019 *Nat. Nanotech.* **14** 35
[20] Wang Z, Joshi S, Savel'ev S E, Jiang H, Midya R, Lin P, Hu M, Ge N, Strachan J P, Li Z, Wu Q, Barnell M, Li G L, Xin H L, Williams R S, Xia Q and Yang J J 2017 *Nat. Mater.* **16** 101
[21] Merolla P A, Arthur J V, Alvarez-Icaza R, Cassidy A S, Sawada J, Akopyan F, Jackson B L, Imam N, Guo C, Nakamura Y, Brezzo B, Vo I, Esser S K, Appuswamy R, Taba B, Amir A, Flickner M D, Risk W P, Manohar R and Modha D S 2014 *Science* **345** 668
[22] Hodgkin A L and Huxley A F 1952 *The Journal of Physiology* **117** 500
[23] Gerstner W, Kistler W M, Naud R and Paninski L 2014 *Neuronal Dynamics: From Single Neurons to Networks and Models of Cognition* (Cambridge: Cambridge University Press) p. 22
[24] Duan Q, Jing Z, Zou X, Wang Y, Yang K, Zhang T, Wu S, Huang R and Yang Y 2020 *Nat. Commun.* **11** 3399
[25] Bortolotto Z A, Amici M, Anderson W W, Isaac J T R and Collingridge G L 2011 *Current Protocols in Neuroscience* **54** 6
[26] Sweatt J D 1999 *Learn. Mem.* **6** 399
[27] Frey U, Krug M, Reymann K G and Matthies H 1988 *Brain Research* **452** 57
[28] Kelleher R J, Govindarajan A and Tonegawa S 2004 *Neuron* **44** 59
[29] Emptage N J, Reid C A, Fine A and Bliss T V P 2003 *Neuron* **38** 797
[30] Hebb D O 2005 *The Organization of Behavior: A Neuropsychological Theory* (Psychology Press) p. 45
[31] Izhikevich E M and Desai N S 2003 *Neural Computation* **15** 1511
[32] Tibbetts P E 2013 *The Quarterly Review of Biology* **88** 139
[33] Bear M, Connors B and Paradiso M A 2020 *Neuroscience: Exploring the Brain, Enhanced Edition: Exploring the Brain, Enhanced Edition* (Jones & Bartlett Learning) p. 54
[34] Zucker R S and Regehr W G 2002 *Annual Review of Physiology* **64** 355
[35] Abbott L F and Regehr W G 2004 *Nature* **431** 796
[36] Bi G and Poo M 2001 *Annual Review of Neuroscience* **24** 139
[37] Tsodyks M V and Markram H 1997 *Proc. Natl. Acad. Sci. USA* **94** 719
[38] Sjöström J and Gerstner W 2010 *Scholarpedia* **5** 1362
[39] Zhang L I, Tao H W, Holt C E, Harris W A and Poo M 1998 *Nature* **395** 37
[40] Allen C B, Celikel T and Feldman D E 2003 *Nat. Neurosci.* **6** 291
[41] Gerstner W, Kempter R, van Hemmen J L and Wagner H 1996 *Nature* **383** 76
[42] Kempter R, Gerstner W and van Hemmen J L 1999 *Phys. Rev. E* **59** 4498
[43] Bohte S M, Kok J N and La Poutré H 2002 *Neurocomputing* **48** 17
[44] Lee J H, Delbruck T and Pfeiffer M 2016 *Frontiers in Neuroscience* **10** 508
[45] Shouval H, Wang S and Wittenberg G 2010 *Frontiers in Computational Neuroscience* **4** 113
[46] Caporale N and Dan Y 2008 *Annual Review of Neuroscience* **31** 25
[47] Cao Y, Chen Y and Khosla D 2015 *Int. J. Comput. Vis.* **113** 54
[48] Diehl P U, Neil D, Binas J, Cook M, Liu S C and Pfeiffer M 2015 *2015 International Joint Conference on Neural Networks*, July 12–17, 2015, Killarney, Ireland p. 1
[49] Sengupta A, Ye Y, Wang R, Liu C and Roy K 2019 *Frontiers in Neuroscience* **13** 95
[50] Zhang X, Wang W, Liu Q, Zhao X, Wei J, Cao R, Yao Z, Zhu X, Zhang F, Lv H, Long S and Liu M 2018 *IEEE Electron Device Lett.* **39** 308
[51] Pickett M D, Medeiros-Ribeiro G and Williams R S 2013 *Nat. Mater.* **12** 114
[52] Han J K, Yun S Y, Lee S W, Yu J M and Choi Y K 2022 *Advanced Functional Materials* **32** 2204102

[53] Zhang Y, Mao G Q, Zhao X, Li Y, Zhang M, Wu Z, Wu W, Sun H, Guo Y, Wang L, Zhang X, Liu Q, Lv H, Xue K H, Xu G, Miao X, Long S and Liu M 2021 *Nat. Commun.* **12** 7232

[54] Rao M, Tang H, Wu J, Song W, Zhang M, Yin W, Zhuo Y, Kiani F, Chen B, Jiang X, Liu H, Chen H Y, Midya R, Ye F, Jiang H, Wang Z, Wu M, Hu M, Wang H, Xia Q, Ge N, Li J and Yang J J 2023 *Nature* **615** 823

[55] Baek K, Park S, Park J, Kim Y M, Hwang H and Oh S H 2017 *Nanoscale* **9** 582

[56] Wang W, Fujita S and Wong S S 2009 *IEEE Electron Device Lett.* **30** 763

[57] Wang W, Fujita S and Wong S S 2009 *IEEE Electron Device Lett.* **30** 733

[58] Inoue I H, Yasuda S, Akinaga H and Takagi H 2008 *Phys. Rev. B* **77** 035105

[59] Cao X, Li X, Gao X, Yu W, Liu X, Zhang Y, Chen L and Cheng X 2009 *J. Appl. Phys.* **106** 073723

[60] Goux L, Lisoni J G, Wang X P, Jurczak M and Wouters D J 2009 *IEEE Transactions on Electron Devices* **56** 2363

[61] Lim E and Ismail R 2015 *Electronics* **4** 586

[62] Yang R 2020 *Chin. Phys. B* **29** 097305

[63] Wang Z, Yin M, Zhang T, Cai Y, Wang Y, Yang Y and Huang R 2016 *Nanoscale* **8** 14015

[64] Huang X, Wu H, Bin Gao, Sekar D C, Dai L, Kellam M, Bronner G, Deng N and Qian H 2016 *Nanotechnology* **27** 395201

[65] Hao Z, Gao B, Xu M, Hu Q, Zhang W, Li X, Sun F, Tang J, Qian H and Wu H 2021 *IEEE Electron Device Lett.* **42** 1276

[66] Zhao M, Gao B, Xi Y, Xu F, Wu H and Qian H 2019 *IEEE Journal of the Electron Devices Society* **7** 1239

[67] Wang X H, Wu H, Gao B, Li X, Deng N and Qian H 2018 *IEEE Electron Device Lett.* **39** 192

[68] Wu W, Wu H, Gao B, Deng N, Yu S and Qian H 2017 *IEEE Electron Device Lett.* **38** 1019

[69] Li Y, Tang J, Gao B, Sun W, Hua Q, Zhang W, Li X, Zhang W, Qian H and Wu H 2020 *Adv. Sci.* **7** 2002251

[70] Wang Z, Rao M, Midya R, Joshi S, Jiang H, Lin P, Song W, Asapu S, Zhuo Y, Li C, Wu H, Xia Q and Yang J J 2018 *Advanced Functional Materials* **28** 1704862

[71] Liu S, Lu N, Zhao X, Xu H, Banerjee W, Lv H, Long S, Li Q, Liu Q and Liu M 2016 *Advanced Materials* **28** 10623

[72] Le Gallo M and Sebastian A 2020 *J. Phys. D: Appl. Phys.* **53** 213002

[73] Kato T and Tanaka K 2005 *Jpn. J. Appl. Phys.* **44** 7340

[74] Pirovano A, Lacaita A L, Benvenuti A, Pellizzer F and Bez R 2004 *IEEE Transactions on Electron Devices* **51** 452

[75] Loke D, Lee T H, Wang W J, Shi L P, Zhao R, Yeo Y C, Chong T C and Elliott S R 2012 *Science* **336** 1566

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

[77] 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

[78] Zhu J, Zhang T, Yang Y and Huang R 2020 *Appl. Phys. Rev.* **7** 011312

[79] Ma J, Ma J, Zhang Q, Peng R, Wang J, Liu C, Wang M, Li N, Chen M, Cheng X, Gao P, Gu L, Chen L Q, Yu P, Zhang J and Nan C W 2018 *Nat. Nanotech.* **13** 947

[80] Luo J, Wu S, Huang Q, Huang R, Yu L, Liu T, Yang M, Fu Z, Liang Z, Chen L, Chen C and Liu S 2019 *2019 IEEE International Electron Devices Meeting (IEDM)*, December 7–11, 2019 San Francisco, CA, USA p. 6

[81] Chanthbouala A, Garcia V, Cherifi R O, Bouzehouane K, Fusil S, Moya X, Xavier S, Yamada H, Deranlot C, Mathur N D, Bibes M, Barthélémy A and Grollier J 2012 *Nat. Mater.* **11** 860

[82] Wang W G, Li M, Hageman S and Chien C L 2012 *Nat. Mater.* **11** 64

[83] Baek S C, Amin V P, Oh Y W, Go G, Lee S J, Lee G H, Kim K J, Stiles M D, Park B G and Lee K J 2018 *Nat. Mater.* **17** 509

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

[85] Ralph D C and Stiles M D 2008 *J. Magn. Magn. Mater.* **320** 1190

[86] Parkin S S P, Kaiser C, Panchula A, Rice P M, Hughes B, Samant M and Yang S H 2004 *Nat. Mater.* **3** 862

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

[88] 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

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

[90] Hsieh H Y and Tang K T 2012 *IEEE Trans. Neural Netw. Learning Syst.* **23** 1065

[91] TermehYousefi A, Azhari S, Khajeh A, Hamidon M N and Tanaka H 2017 *Materials Science and Engineering: C* **77** 1098

[92] Kim Y, Chortos A, Xu W, Liu Y, Oh J Y, Son D, Kang J, Foudeh A M, Zhu C, Lee Y, Niu S, Liu J, Pfattner R, Bao Z and Lee T W 2018 *Science* **360** 998

[93] Wang W, Pedretti G, Milo V, Carboni R, Calderoni A, Ramaswamy N, Spinelli A S and Ielmini D 2018 *Sci. Adv.* **4** eaat4752

[94] Wang D, Shan H and Xin J 2018 *Hearing Research* **370** 283

[95] Tan H, Zhou Y, Tao Q, Rosen J and Van Dijken S 2021 *Nat. Commun.* **12** 1120

[96] Meng J, Wang T, Zhu H, Ji L, Bao W, Zhou P, Chen L, Sun Q Q and Zhang D W 2022 *Nano Lett.* **22** 81

[97] Li C, Hu M, Li Y, Jiang H, Ge N, Montgomery E, Zhang J, Song W, Dávila N, Graves C E, Li Z, Strachan J P, Lin P, Wang Z, Barnell M, Wu Q, Williams R S, Yang J J and Xia Q 2018 *Nat. Electron.* **1** 52

[98] Chu M, Kim B, Park S, Hwang H, Jeon M, Lee B H and Lee B G 2015 *IEEE Transactions on Industrial Electronics* **62** 2410

[99] Valenzuela W, Saavedra A, Zarkesh-Ha P and Figueroa M 2022 *Sensors* **22** 6538

[100] Nose A, Yamazaki T, Katayama H, Uehara S, Kobayashi M, Shida S, Odahara M, Takamiya K, Matsumoto S, Miyashita L, Watanabe Y, Izawa T, Muramatsu Y, Nitta Y and Ishikawa M 2018 *Sensors* **18** 1313

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

[102] Hirata T, Murata H, Arii T, Matsuda H, Yonemochi H, Tezuka Y and Tsunai S 2022 *IEEE Open Journal of Circuits and Systems* **3** 311

[103] Wang F, Hu F, Dai M, Zhu S, Sun F, Duan R, Wang C, Han J, Deng W, Chen W, Ye M, Han S, Qiang B, Jin Y, Chua Y, Chi N, Yu S, Nam D, Chae S H, Liu Z and Wang Q J 2023 *Nat. Commun.* **14** 1938

[104] Yu R, He L, Gao C, Zhang X, Li E, Guo T, Li W and Chen H 2022 *Nat. Commun.* **13** 7019

[105] Lee D, Park M, Baek Y, Bae B, Heo J and Lee K 2022 *Nat. Commun.* **13** 5223

[106] Seo S, Jo S H, Kim S, Shim J, Oh S, Kim J H, Heo K, Choi J W, Choi C, Oh S, Kuzum D, Wong H S P and Park J H 2018 *Nat. Commun.* **9** 5106

[107] Tan H and van Dijken S 2023 *Nat. Commun.* **14** 2169

[108] Yin L, Han C, Zhang Q, Ni Z, Zhao S, Wang K, Li D, Xu M, Wu H, Pi X and Yang D 2019 *Nano Energy* **63** 103859

[109] Wang H, Zhao Q, Ni Z, Li Q, Liu H, Yang Y, Wang L, Ran Y, Guo Y, Hu W and Liu Y 2018 *Advanced Materials* **30** 1803961

[110] Gao S, Liu G, Yang H, Hu C, Chen Q, Gong G, Xue W, Yi X, Shang J and Li R W 2019 *ACS Nano* **13** 2634

[111] Jaafar A H, Gray R J, Verrelli E, O'Neill M, Kelly S M and Kemp N T 2017 *Nanoscale* **9** 17091

[112] Tan H, Ni Z, Peng W, Du S, Liu X, Zhao S, Li W, Ye Z, Xu M, Xu Y, Pi X and Yang D 2018 *Nano Energy* **52** 422

[113] Eginligil M, Cao B, Wang Z, Shen X, Cong C, Shang J, Soci C and Yu T 2015 *Nat. Commun.* **6** 7636

[114] Choi W, Cho M Y, Konar A, Lee J H, Cha G-B, Hong S C, Kim S, Kim J, Jena D, Joo J and Kim S 2012 *Advanced Materials* **24** 5832

[115] Zhou F, Zhou Z, Chen J, Choy T H, Wang J, Zhang N, Lin Z, Yu S, Kang J, Wong H-S P and Chai Y 2019 *Nat. Nanotechnol.* **14** 776

[116] Zhang Z, Zhao X, Zhang X, Hou X, Ma X, Tang S, Zhang Y, Xu G, Liu Q and Long S 2022 *Nat. Commun.* **13** 6590

[117] Wu N 2018 *Sci. China Inf. Sci.* **61** 060421

[118] Jing C O, Zhao Z, Nan Q I, Liu L and Wu N J 2019 *Acta Photonica Sinica* **48** 0704001

[119] Ara Shawkat M S, Sayyarparaju S, McFarlane N and Rose G S 2020 *2020 IEEE 63rd International Midwest Symposium on Circuits and Systems*, August 9–12, 2020, Springfield, MA, USA, p. 377