IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 13 March 2024, accepted 7 April 2024, date of publication 10 April 2024, date of current version 22 April 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3387094

TOPICAL REVIEW

FETs for Analog Neural MACs

RINKU RANI DAS, T. R. RAJALEKSHMI, SRUTHI PALLATHUVALAPPIL, (Graduate Student Member, IEEE), AND ALEX JAMES, (Senior Member, IEEE)
School of Electronic Systems and Automation, Digital University Kerala, Thiruvananthapuram 695317, India

Corresponding author: Alex James (apj@ieee.org)

This work was supported by the Ministry of Electronics and IT, Government of India.

ABSTRACT This study provides a comprehensive view on neural network systems with implemented with crossbar circuits, and device-level understanding of modern FET technologies in neuromorphic computing. This work categorizes and analyzes various transistor types, including ion-gate, ferroelectric, and floating-gate transistors, shedding light on their unique advantages and applications in neuromorphic computing. In this overview, we explore the fundamental principles, recent advancements, and significant trends in transistor-based neuromorphic devices, providing valuable insights into this innovative field. This work also examines resistive memories and 2D materials, that could revolutionize transistor fabrication for neuromorphic devices. Further, various research challenges, limitations, and potential research directions are discussed.

INDEX TERMS Field effect transistor (FET), CMOS, memristors, multiply and accumulate (MAC), 2D materials, FeFET, IGT.

I. INTRODUCTION

The wide range of hardware implementations in the analog, digital or mixed signal domains of bio-inspired computing systems contributes to the era of neuromorphic computing. Neuromorphic computing closely resides in analog computing hardware similar to the human brain. The expected functionalities and how closely they mimic the human brain are the most important factors in neuromorphic computing hardware. It attempts to integrate various biological characteristics of the human brain such as synapses, synaptic weight, synaptic transmission, spike-time encoding, different types of plasticity, parallel processing of accumulated information and higher cognitive functions. The extremely low power consumption, fast processing of information and high density of the human brain are the leading factors on which the evolving neural network implementations in the digital, analog and mixed-signal domains are the focus. Extreme energy efficiency is accounted for in advanced spiking neural networks. The analogies between biological neural networks and hardware systems can implement using transistors by making them operate in a region that follows the dynamics of neurons and synaptic dynamics. The amount of data being generated, processed and stored is increasing, and the higher energy consumption required for the same is due to the data traffic between memory and processing, which is separated in the Von-Neumann architecture. This causes huge energy wastage and reduces computational speed. The possibility of integrating memory and processing units taking inspiration from the human brain via neuromorphic chips addresses the bottleneck for data transmission, which helps to acquire energy consumption and computational speed similar to that of the dynamics of the human brain. One key challenge for neuromorphic computing hardware implementation is the different non-idealities and variabilities of memory devices along with transistor technology. To accelerate more than conventional digital circuits, the complexity of building neuromorphic chips that require analog behavior of memory devices and transistors needs to be addressed further. The lack of hierarchical boundaries, which von-Neumann allows for development, is lacking in the case of neuromorphic computing and needs to be addressed to accelerate its evolution [1]. While building neuromorphic chips, many neurons and synapses must be modelled to make them computationally powerful and to imitate brain functions such as non-linearity, spiking behavior, plasticity and long-term

The associate editor coordinating the review of this manuscript and approving it for publication was Chaitanya U. Kshirsagar.

memory. Millions of biological interconnections, such as that of the human brain, need to be replicated in neuromorphic chips to mimic high degrees of interconnection.

Analog computing, in-memory computing, spike coding, task-specific connectivity, and parallel processing are the key tricks employed by the human brain that can be transferred to neuromorphic computing hardware. Currently, it is difficult to mimic all the features of the human brain, but it can be made comparable by starting from conventional technologies and then exploring new dimensions. Neuromorphic computing is progressing to cover the information storage and processing principles of the human brain. For example, the scientific community has already performed sound, image and video processing for recognition, which is mostly at the theoretical and algorithmic levels using computers based on the Von-Neumann architecture. Currently, researchers are putting great effort into implementing the functions using novel architectures in a manner where the human nervous system performs storage and computations that range from single-memory devices to in-memory computing architectures. To build efficient neural computing architectures that perform beyond Von-Neumann's computing architecture, scientists must bridge the gap between today's electronic chips and human intelligence. The characteristics of the human brain, such as storing information in the analog domain and plastic rather than having limited reconfigurability, million synapses and neurons, and multiple dynamics rather than having a single clock, are to name a few, taking into account the evolving neuromorphic computing hardware design [2]. The architectural and computational principles of the human brain, with an energy consumption of approximately 20 Watts, need to be replicated in neural chips, which is a crucial challenge to address. Different features of the human brain still cannot be understood like laminar architecture, diversity (for example, various types of distinguishable neuron cells exist, but most deep neural networks use identical neurons), a high degree of organization (specific areas for specific tasks, which are usually not considered in neural networks) and synapses that have complicated biochemical processes encoded as a single weight value in neural networks. Bridging these gaps opens the possibility of well-established neuromorphic computing architectures that have the potential to store and compute similar to the human brain. Neuromorphic hardware is biologically inspired by computational aspects of the human brain. Mapping between biological and artificial computations requires a set of in-memory computing circuitries [3]. Neural networks can be implemented using transistor-based resistive memory circuits. The Weighted summation operations, succeeding activation functions, and remaining intermediate operations can be implemented using specific circuits such as one-transistor-one-memristor-based crossbars, opamp-based analog-to-digital converters, sample and hold circuits, purely transistor-based control switches, activation functions and transmission gates. The memristors are artificial synapses. Voltage pulses are applied to change the conductance states of the memristor to which the input weights of the neural networks are mapped, which is analogous with the application of neural spikes are to change the synaptic weights [4], [5], [6]. The memristive device and selector transistors are arranged in rows and columns to implement weighted summation operations [7]. Other transistor technology-based circuits are used for further computations to implement neural networks.

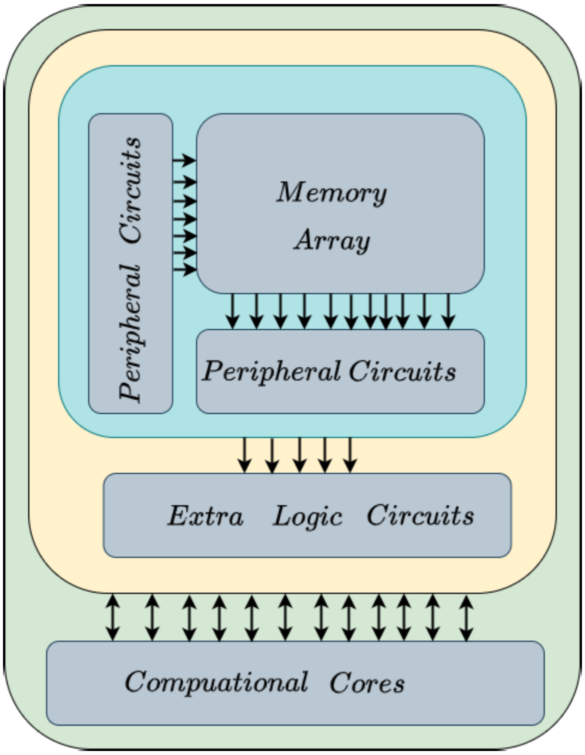

FIGURE 1. The system-level architecture blocks of in-memory computing.

The most significant functional units of neural systems are the neurons and synapses [8]. Scientists and engineers can create artificial systems, by understanding and replicating the behavior of these components. Next-generation neuromorphic computing that mimics the human brain, using advanced techniques to boost computational power and pattern recognition, is receiving significant attention [9]. Neurons receive information from sensory organs or other neurons. It then generates an electrical signal after a certain threshold point called the action potential. Finally, the electrical signal is transmitted over long distances within the nervous system. Synapses transmit signals from one neuron to another. Neurons receive signals from thousands of neurons via synapses. Synapses act as neural bridges where information is stored and processed through dynamic adjustments in their connection strengths, enabling the brain to learn and form memories [10]. This neural system can integrate seamlessly with a parallel in-memory computing architecture, offering low power consumption and enabling more precise information processing [11]. Traditional digital computing is performed using CMOS technology to simulate biological synapses and neurons [12]. However, digital computing using conventional CMOS technology has various unwanted issues [13]. It is very difficult to handle the information between physically separated computing and memory units. Conventional computing consumes more power, speed, and work overheads. To overcome these limitations various types of modified transistors has been proposed and studied. Electrolyte-gated transistors (EGTs), ferroelectric-gate field effect transistor (FET) (FeFETs), and floating Gate (FG) FET have been recognized as highly promising neuromorphic devices for emulating both neurons and synapses [14]. Their unique characteristics help to mimic the behaviour of synapses and neurons more closely than traditional digital computation units based on complementary metal oxide semiconductor (CMOS) technology. By examining the interplay between neural network architectures and transistor-based devices, this comprehensive review focuses on the pivotal role of transistors in realizing neuromorphic computing.

This study encompasses a wide spectrum of technical details, spanning from modular and three-dimensional crossbar arrays to circuit-level implementations and transistor technologies. It focuses on covering the core concepts of neuromorphic computing and examining its fundamental principles alongside neural network architectures, three-dimensional crossbar arrays, and circuit-level implementations. We provide concise progress in transistor technology over time and explore the latest advancements in transistor-based neuromorphic devices. We discuss the working mechanism of transistors, followed by their integration in neuromorphic computing. Additionally, we conducted an in-depth analysis of the materials crucial to these transistors, emphasizing their significance in advancing the field. The focus is on materials such as silicon, ferroelectric materials, and 2D materials, which have the potential to revolutionize the fabrication of transistors for neuromorphic devices. As neuromorphic computing continues to grow, it is crucial to understand how transistors and neural networks work together to improve cognitive computing. The review concludes by addressing the challenges faced in this field and discussing future perspectives.

II. NEUROMORPHIC COMPUTING

Fig. 1 shows Computation-In-Memory(CIM), in which computation occurs in the memory core. The system-level architecture integrates processing within memory. The architecture mainly consists of non-volatile memory devices that enable storage and computation. The memory core comprises a memory array and a peripheral circuit. The computational results are produced within the array or in the periphery. Based on where the computational results are formed, the CIM architecture can be classified into CIM-Array and CIM-Periphery. The computational results are produced and stored in the form of resistance states within the memory array in the case of CIM-Array. Because computation and storage occur within the same memory array, the maximum bandwidth can be acquired for transferring data between the computation and memory. High parallelism can be achieved because the computations are performed independently of sense amplifiers. Endurance and energy issues occur because of the frequent write operations. In complex functionalities, there is a chance of performance overhead owing to device programming and cascading. The CIM-Array requires significant design efforts.

In CIM-Periphery, computational results are produced within the periphery circuitry. Memory periphery circuits are mostly based on CMOS technology; the output is voltage. Because the memory states do not change during computation or post-computation, this type of architecture will not affect the endurance of the memory array. During computations, the sense amplifiers and analog-to-digital converters are shared, which causes performance degradation.

The Von-Neumann architecture faces various challenges while scaling down existing technology without compromising efficient computation. The complexity and large volume requirement for realizing neural network algorithms to hardware demands the characteristics of scalability, non-volatility, high density for integration, and low area and power consumption. The evolving resistive memory devices and memristors integrated with transistor technology are good candidates for mimicking synapses in hardware neural network implementations. Memristive devices mimic synapses in the human brain. Hence, memristors mimic synapses in neural network hardware implementations. The characteristics of memory resistors such as scalability, non-volatility, transistor compatibility, high density for integration, and low power and area consumption, make them suitable for neural network implementations. They are also flexible for acting as multi-level cell memory, Scaling by three-dimensional stacking, and multilayer cells, which gives a higher level of scalability. Neural network weights are mapped to the conductance values of the memristor, which helps to emulate the weighted summation operation in neural network implementation.

The emerging Neuromorphic computing [15] technology brings high-performance computations in analog, digital and mixed-signal domains that are brain-inspired [16]. Neuromorphic computers draw inspiration from the structure and functions of neurons and synapses found in the human brain. Processing and storage functionalities highly resemble human intelligence. Neuromorphic computers use programming algorithms based on different neural networks. It differs from Von-Neumann computers. The processing and storing are performed by the CPU and memory units separately in the case of Von-Neumann. Hence, neuromorphic computing simplifies computations through parallel processing and storage without requiring separate locations.

The computations performed in neuromorphic computing architectures are parallel and comparatively simple compared to the Von-Neumann architecture [17]. The collocation of memory and processing helps reduce the throughput and power requirements required for data transfer. The flexibility

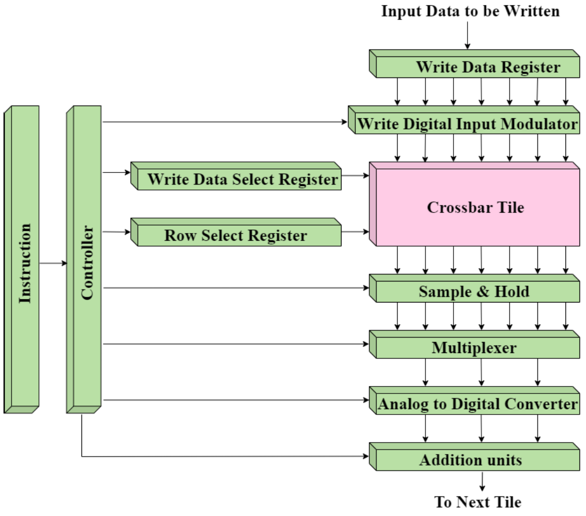

FIGURE 2. Functional blocks of neuromorphic computing hardware architecture.

to scale helps build larger networks based on different applications in artificial intelligence, machine learning, etc.

The Spiking Neural Network (SNN) is inspired by biological neural systems and performs computations in which neurons and synapses include notions of time. SNN algorithms comprise neurons and synapses associated with time delay for neuromorphic computing, which is significantly different from other neural networks [18]. Different neuron models are there like Integrate-and-Fire, Hodgkin-Huxley neuron model, etc. Spikes fire when the charge is integrated over time and meets the threshold in the case of Integrate-and-Fire. The neurons will have a rest time after firing, which is called the refractory period. In SNN, both neurons and synapses include time components based on which functionalities vary. Based on the network activity, the synapse's weight value changes. In a neuromorphic architecture, one can program neuron thresholds, delay values, and synaptic weights. In realizing SNN on neuromorphic hardware, information asynchronously propagates throughout the network, which can be treated as event-driven and fits well with temporal dynamics.

For Spiking Neural Networks, TENNLab has different architectures. Two of these are WHETSTONE, NIDA, MrDANNA and DANNA2. The DANNA2 is a two-dimensional architecture. Properties such as synaptic weight and neuron threshold are represented by integers. In NIDA, the length of the synapses determines synaptic delays in the three-dimensional architecture. In MrDANNA, the architecture was implemented using memristors. The three-dimensional spiking architecture WHETSTONE stores elements at a floating point. TENNLab implemented a two-dimensional spiking architecture with FPGA, software and Very Large Scale Integration implementations [19].

The modulation of synaptic strength is contingent on the activity of connected neurons, a characteristic that can be theorized as a learning mechanism. Spike-timing-dependent plasticity (STDP) is the prevalent synaptic plasticity mechanism employed in neuromorphic computing. It involves adjusting weights based on the relative spike timings between pre and post-synaptic neurons. Recurrent Neural Networks with synaptic plasticity and delays are a broader class of SNN used for modelling. An example of such a class of networks is polychronization networks, which are implemented for spatio-temporal classifications.

A. NEUROMORPHIC COMPUTING: ARCHITECTURES

AI systems consistently outperform computation and storage. The in-memory computing approach which closely resides in the human brain is inspired by the learning, processing, and storing capabilities of the human brain with extremely low power, latency, and highly dense architecture. Memristive devices that offer highly dense memories can potentially emulate the human brain. They are analog programmable devices that can be programmed to desired conductance states, and they retain their latest attained resistance value. It acts as a non-volatile memory. Because computations are naturally analog, analog memristive computing is widely accepted and considers memory and the connection of memories as a type of intelligence. The possibility of lower signal attenuation issues, reduced parasitic impact, computation process before noise build-up, performance of Multiply and Accumulate (MAC)operation, reduction in the number of computational blocks, and dense architecture with reduced power, area, and latency enhances the demand for memristive computing in neuromorphic applications [20], [21], [22], [23].

In the crossbar arrangement, memristors and transistors are arranged in a matrix form. Transistors are selector devices. Each memory cell consists of a memristor and transistor [24]. There are multiple inputs and outputs. The Input voltages are fed through the rows and the output currents are read from the columns. The current read equals the result of the multiplication and accumulation operation carried out between the input voltages and the equivalent conductance of the memristor and transistor. The Multiplication and accumulation operations in a memristive crossbar emulate the weighted summation operation in neural network [25], [26]. In neural network implementations, the weights are mapped to the conductance values of the memristor. Some of the application areas of crossbars are neural networks (Artificial Neural Networks, Convolutional Neural Networks, Deep Neural Networks, Spiking Neural Networks, Cellular Neural Networks), Analog/Digital Memory (associative, long-term memory, Multi-level memory, NAND.etc), and Solvers (Linear Equations, Partial Differential Equations, Markov Chains,.etc), analog/digital logic gates (threshold logic, bio-inspired,.etc), cryptography (PUF), image processing (Cellular Neural Networks, object detection, edge detection, face detection, etc.).

Fig. 2 shows a detailed neuromorphic architecture block diagram having the major components along with the

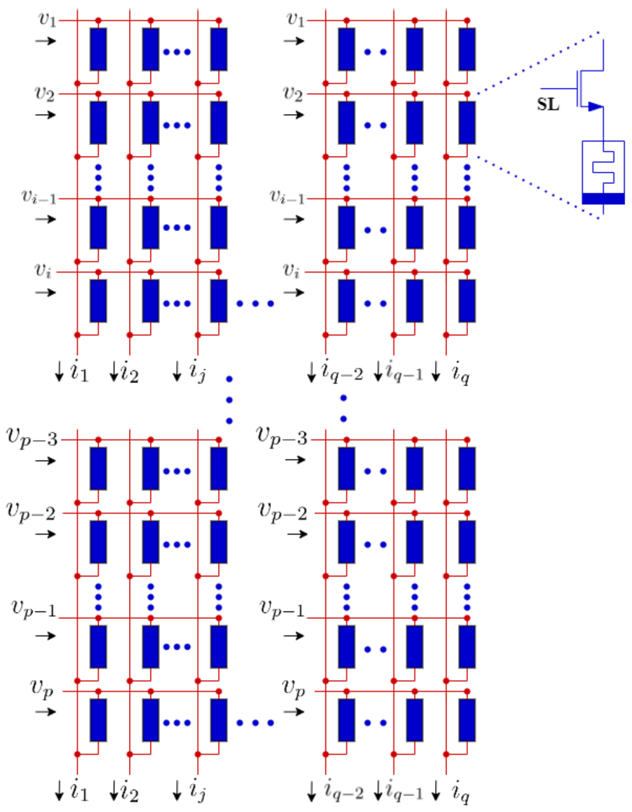

FIGURE 3. The tiled architecture of a memristive crossbar with $p \times q$ cells implemented using multiple sub crossbars each of size $i \times j$. The number of sub-crossbars for rows and columns is c and d where $p = c \times i$ and $q = d \times j$.

crossbar array. The write, read, and computation operations are executed using a crossbar. Computational operations include logical, addition, and multiplication operations.

1) WRITE OPERATION

To which row and column of the crossbar the data need to be fed should be specified. Three registers are assigned to perfom the task. The Write Data register (WD register) is used to feed the data, that be written to the crossbar. The length of the WD register depends on the width of the crossbar and the number of conductance levels that the memristors can program. To address the issues of endurance and power consumption, it is not necessary to select all elements in the array for writing data. A Write Data Select register (WDS register) is used to select the columns to be activated (Particularly for write-verify operation). A Row Select register (RS register) is used to select the desired row to which the data must be written. Depending on the crossbar technology, the voltage required to be applied to the crossbar varies. Digital-to-Analog Converter (DAC) converts data from digital to analog. Different voltage levels must be applied to the gate and source of the target row.

2) READ AND COMPUTATIONAL OPERATIONS

In architecture, the output generated by different operations must be read by the peripheral circuit. This can be a direct memory read or the result of computational operations. After matrix multiplication, the generated analog result needs to be captured by a sample and hold circuit. This sample and hold circuit helps to separate the execution in an array and the operations in read-out circuitry. This helps to pipeline the system. The results, in analog form, are given to the Analog-to- Digital Converter/ Sense Amplifiers (ADC/SA) for conversion to the digital domain. To avoid issues of high area and power consumption, the ADCs are not allocated to each column. Multiplexers are used to share several columns with an ADC. Additional computational operations were performed using the addition unit.

III. MODULAR AND THREE-DIMENSIONAL CROSSBAR ARRAY ARCHITECTURES

The neural network implementation of a memristive crossbar array uses two architectures, two-dimensional and three-dimensional. The two-dimensional tiled architecture of a memristive crossbar array is more widely used than the three-dimensional architecture, which is still evolving.

A. MODULAR ARCHITECTURE

To achieve highly complex neural applications, a large memristive crossbar array must be created. Large memristive crossbar arrays are limited by the sneak current issue, which causes read-out current errors, a lack of accuracy, and power loss. The influence of the sneak path issues is addressed by dividing the larger crossbar array into smaller ones as shown in Fig. 3. This modular memristive crossbar array approach helps reduce the IR drop (Intermediate Resistance Drop) and sneak current issues to an extent, especially when it needs to be scaled. Each layer of the memristive crossbar array consists of memristors and selector devices arranged in matrix form. Each node can be accessed using its corresponding rows and columns. These resistive memory devices, along with selector devices, are responsible for emulating the synapses in neural networks. By adjusting the different parameters of these memristive devices through amplitude and frequency adjustment of the applied voltage, with the help of selector devices, different conductance states are attained. A modular crossbar array is designed by breaking a large memristive crossbar array into smaller modules by dividing them into rows from several small modules of equal size. The number of rows in each module equals the total number of rows in the larger array divided by the total number of modules. For each module, the number of columns is the same for each module. The total output current from the column of the larger array is equal to the summation of the currents from the corresponding columns of each module. The Other columns operate simultaneously in the same routing mechanism. This approach provides an appropriate restriction on the path of the leakage current. In the Modular Memristive Crossbar Array approach, the path of the leakage current is restricted by dividing a larger crossbar array into several modules. Owing to the reduced sneak path current, the read-out current error is also reduced.

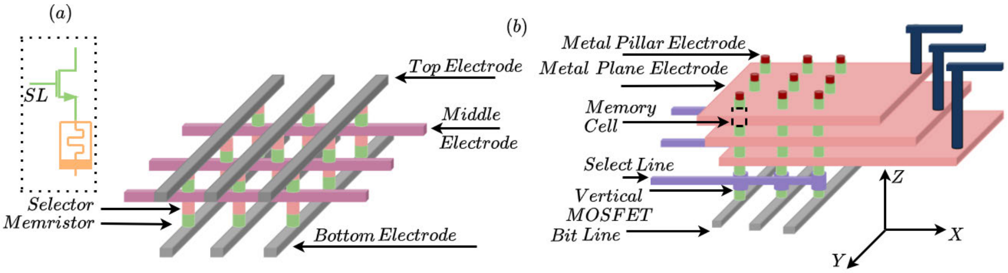

FIGURE 4. Three-dimensional architecture of memristor crossbar array in which each node contains the memory device and transistor (a) Horizontally stacked crossbar array (b) Vertical crossbar array.

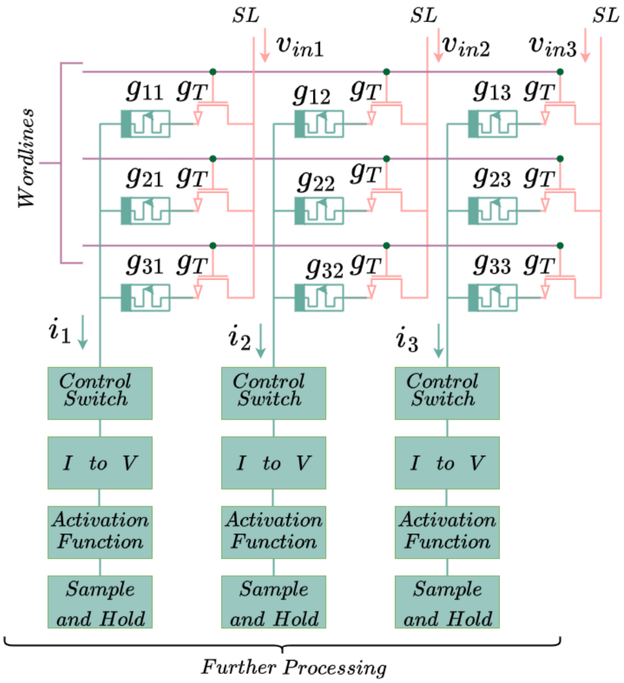

FIGURE 5. Memristor-crossbar architecture with inputs v_1, v_2, v_3 and output currents i_1, i_2,i_3. The conductance of memristors is denoted as g_{mn} and that of transistors is denoted by g_T. The intermediate circuitry consists of a control switch, a current-to-voltage converter, a sample and hold circuit, and circuitry corresponding to the activation function.

A more accurate implementation of neural networks can be achieved because of the reduced relative current error.

B. THREE-DIMENSIONAL ARCHITECTURE

The two-dimensional tiled architecture of a memristive crossbar array restricts the possibility of implementing a neural network with several layers above a specific limit. To address this limitation, the three-dimensional arrangement of crossbar arrays provides a highly dense arrangement of devices. A neural network with many layers does not contribute significantly to the chip area when using a three-dimensional architecture. Reduced latency for operation also helps reduce power. A more advanced scheduling mechanism for writing, reading, and computation operations in a three-dimensional architecture further helps reduce energy consumption.

When hardware neuromorphic computing requires wider and deeper neural networks to implement complex functionalities, the three-dimensional integration of memristive crossbar arrays will be more efficient and effective, as shown in Fig. 4. Hardware implementation of a neuromorphic chip for high-density applications, memristors can be scaled with three-dimensional integration to function as multilayered neural networks with minimum area requirement. Three-dimensional integration can be performed in two ways. Vertically stacked three-dimensional crossbar arrays and horizontally stacked three-dimensional crossbar arrays [27].

In the three-dimensional horizontally stacked crossbar arrays shown in Fig. 4(a), a higher density can be attained by, vertically stacking two-dimensional crossbar arrays. Owing to the flexibility to scale laterally, peripheral circuits can be placed under crossbar arrays to obtain a more compact design, and a separate selector or transistor can be accompanied by a resistive memory. Hence, the selector and transistor can be individually optimized. The number of interconnections can be reduced to simplify the fabrication process using a shared middle electrode. This shared middle electrode can be either a bit line or word line. If bit lines are shared, the number of peripheral circuits can be reduced because the connected read-out circuits, including sense amplifiers and current-to-voltage converters, can be reduced. Hence, it is preferable to obtain higher power and area efficiency. In this type of three-dimensional integration, a minimum of three major lithography and etching steps are required for fabrication. The challenge in the three-dimensional integration is increasing the number of stacking layers, and the number of interconnects also increases accordingly and demands more lithographic and etching processes. Staircasing the interconnects to both the bit lines and word lines will be another challenge, because different lengths of the interconnections cause different voltage drops among different layers.

By fabricating vertical crossbar arrays, as shown in Fig. 4(b), the challenges encountered with horizontal crossbar arrays can be addressed. This can be treated as a word-plane-type crossbar array. Here the plane electrode serves

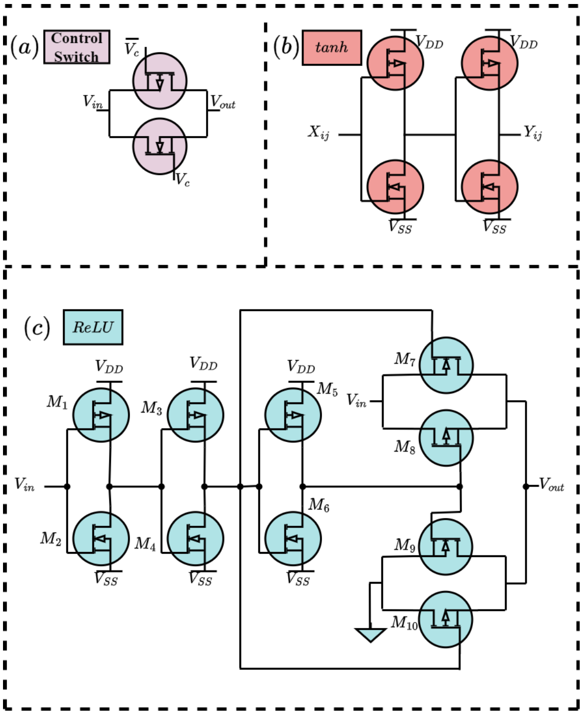

FIGURE 6. (a) Control Switch, (b) Hyperbolic tangent activation function (c) ReLU activation function.

as the wordline. Memristors that can be treated as vertical memory elements are formed at the crosspoints of each pillar electrode and plane electrode. To select the bitline, a vertical mosfet is used in which the gate is controlled by the selected line. In this type of three-dimensional arrangement, only one critical photolithography is needed, and hence the fabrication cost is reduced. In this case, there are also challenges such as etching deep holes.

IV. THE CIRCUIT-LEVEL IMPLEMENTATION OF NEUROMORPHIC COMPUTING ARCHITECTURES

Complex neural networks are integrated with edge devices using CMOS-Memristive neuromorphic circuits because they are programmable, non-volatile devices capable of in-memory computing [28]. This helps to avoid sending large amounts of data collected by edge devices to the processing unit and memory and provides limited area and power consumption. As mentioned in the previous sections, crossbar array nodes consisting of transistor-memristor pairs are used to accelerate multiplication and accumulation operations for neural networks in hardware implementations. Different neural networks such as Artificial Neural Networks, Convolutional Neural Networks, Spiking Neural Networks, and Long Short-Term Memory networks, are implemented in hardware using one transistor-one memristor crossbars.

Fig. 5, 6, and 7 refer to the circuits involved in the neuromorphic hardware implementations. The crossbar circuits, activation functions, control switches, sample and hold, and current-to-voltage converters are based on CMOS circuits. The convolution and deconvolution operations required for neural networks are realized using memristive crossbars and, their sizes depends on the input image. Memristors are programmed to the desired conductance states with the help of transistor selectors based on the mapping of weights; accordingly, a weighted summation is performed. In neural network implementation using a memristive crossbar architecture, control switches control the sequential processing of the rows and columns of the crossbar. The control switches assist in facilitating the sequential processing of the crossbar columns. The switching transistors, accompanied by resistive memory devices, are connected to the control voltage at the drain rather than at the source to improve the linearity of the switch. It also helps reduce the leakage current when the control switch is off. A simple control switch is shown in Fig. 6(a). When the voltage at V_c is elevated, the input to the NMOS is in a high state, while the input to the PMOS is in a low state. Hence, the NMOS and PMOS are turned on. It acts as a common resistance path; in other words, it acts as a short circuit. Input v_{in} is passed to the output. When Vc is low, the input to the NMOS is low, and the PMOS is high. Hence NMOS and PMOS are turned off. It will act as a high resistance path, or in other words, as an open circuit. The input v_{in} will not be passed to the output, and the realization of some activation functions using transistor technology is illustrated in Fig. 6(b) and (c). Fig. 6(b) shows the Hyperbolic tangent (tanh) activation function. The two cascaded inverters and biasing voltages V_{DD} and V_{SS} realize the tanh activation function. Fig. 6(c) shows the activation function ReLU (Rectified Linear Unit). Two transmission gates exist in a ReLU circuit. M_1, M_2, M_7 and M_8 are supplied with input voltage. The threshold voltages of the inverters are 0. When the input signal is positive, M7 and M8 are on, which can be treated as under the condition of the transmission gate, and the input is passed to the output. When the input signal is negative, M9 and M10 are on, which can be treated as the on condition of the second transmission gate, and the input is passed to the output(which is grounded).

The circuit diagram of the opamp and the circuits designed using the opamp are shown in Fig. 7. A circuit diagram of the opamp is shown in Fig. 7(c). It consists of two differential stages and a common gain stage. The output current read from the columns was converted into voltage using opamp-based current-to-voltage converters. As shown in the circuit diagram in Fig. 7(a), the I- to-V converter gives an inverted output after the first stage, which is equal to the product of the input voltage and feedback resistance. The second stage gives a non-inverted output, which also equals the product of the input (output of the first stage) and the feedback resistance. During the sequential processing of crossbar columns, an analog Sample and Hold circuit are used to retain the voltage signal as shown in Fig. 7(b). It comprises two voltage buffers to mitigate the impact of signals from other circuit components on the sampled signal, along with a single sample and hold component. The input is sampled when V_{clk} is high. The output from the activation functions can be read V_{clk} will be high until the last column is read.

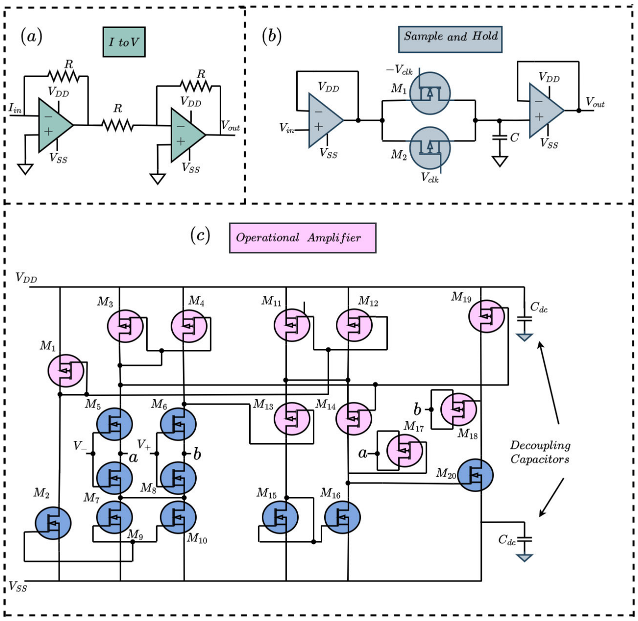

FIGURE 7. (a) Current-to-voltage converter, (b) Sample and hold circuit, (c) Opamp circuit with two differential stages and one gain stage.

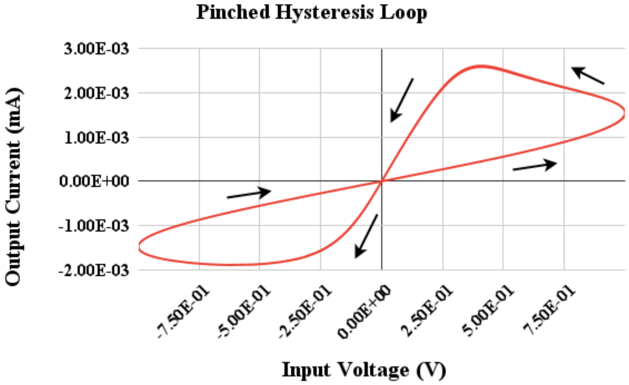

FIGURE 8. Pinched Hysteresis of a memristor in which the output current lags behind the input voltage.

V. MEMRISTIVE DEVICES

Memristors, operating as two-terminal devices, display resistance that is affected by factors such as the magnitude, applied voltage, and polarity. Even in the absence of an applied voltage, the resistance remains, resulting in non-linear and memory characteristics. Several types of memristors exist, each with unique characteristics and materials. Several emerging memristors, including resistive random access memory (ReRAM), Phase Change Memory (PCM), and Spin-Torque Transfer RAM (STT-RAM), operate based on resistive switching in memristor materials. Nevertheless, the specific mechanism of resistive switching varies among memristors.

The characteristics of a memristor showing the variation in the output current with respect to the input voltage is shown in Fig. 8. The output current lags behind the input voltage, resulting in a pinched hysteresis loop that passes through the origin. Each slope refers to a conductance state that can be programmed. When the frequency increased, the area of the pinched hysteresis loop decreased. At a higher frequency, it appears as a straight line.

A. RESISTIVE RANDOM ACCESS MEMORY (RERAM)

ReRAM, which is an emerging resistive memory, has low writing energy and high density. ReRAM is also suitable for building low-latency memories. It also provides a high endurance (10^{10}) also. Technologies that use variations in resistance to store information are called resistive memory. ReRAM especially points to metal-oxide Resistive Random Access Memory because metal oxide is used as the storage medium. ReRAM consists of two electrodes (top and bottom electrodes) and a metal-oxide layer sandwiched between them, as shown in Fig. 9. To switch the resistive state of a ReRAM cell, an external voltage of a particular polarity, duration, and magnitude is applied. SET (switching from high resistance state to low resistance state) and RESET (switching from low resistance state to high resistance state) are controlled by the external voltage. The switching process of the ReRAM is based on the formation and rupture of the conductive filament between the electrodes. The SET process involves the regeneration of conductive filament by drifting oxide ions to the anode (positive electrode) and leaving oxide vacancies in the metal oxide layer. In the RESET process, oxide ions are returned to the oxide layer by the force of an electric field, followed by recombination with oxide vacancies. As a result, the conductive filament is cut off, and the ReRAM is transferred to a high resistance state. The resistance of a memristor can be programmed to vary values between these two high resistance state and low resistance state, by applying voltage pulses and changing their amplitude and frequency. The larger the size of the conductive filament, the smaller the resistance. Multiple states are achieved by changing the strength of the conductive filament, which depends on the applied voltage. In digital memristive devices, the initial formation process of the conductive filament is termed electroforming. After the electroforming process, the filamentary switching model functions. The voltage required for electroforming is higher than that required for switching. Analog memristive devices are free of electroforming. The Characteristics of ReRAM vary with the materials. ReRAM has the advantages of high endurance (10^{10}), scalability, and switching speed with relatively less energy consumption and latency.

B. PHASE CHANGE MEMORY (PCM)

Phase Change Memory (PCM), also known as Perfect RAM (PRAM), PCRAM, and Chalcogenide RAM (CRAM), belongs to the category of emerging non-volatile memories that rely on the principles of chalcogenide glass, and possess two distinct phase states [29]. PCM's switching operation of

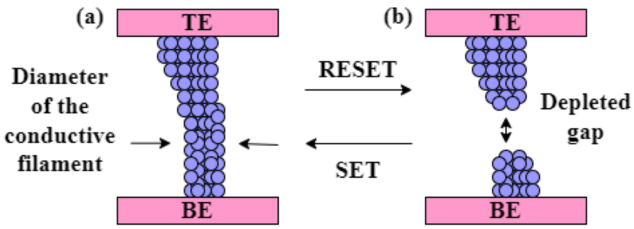

FIGURE 9. ReRAM device: Set (Transition from High Resistance State to Low Resistance State) and Reset (Low Resistance State to High Resistance State) operations. (a) The top Electrode and Bottom Electrode are connected by a Conductive Filament in Low Resistance State. (b) In High resistance, this Conductive Filament disconnects.

PCMs relies on the phase change of the material, transitioning from an amorphous state to a crystalline state. This process involves two distinct resistance levels: low and high. The PCM technology stores information through the transition from a low-resistance crystalline state to a high-resistance amorphous state. Shifting from the amorphous phase to the crystalline phase is considered the SET process whereas the reverse method from the crystalline phase to the amorphous phase is considered RESET switching. The PCM cell architecture is presented in Fig.10 (a). PCM cells exhibit a low resistance state (LRS) at high temperatures. Applying an external power supply and current allows PCM cells to rapidly RESET to a high-resistance state (HRS) in a short time-period. This RESET process involves a transition from a crystalline state to amorphous phase state. Achieving a return to a crystalline state involves the SET process, which requires an external current pulse at the melting temperature. However, the crystallization process requires a longer duration. The I-V characteristics of the PCM cells are presented in Fig. 10 (b) [30].

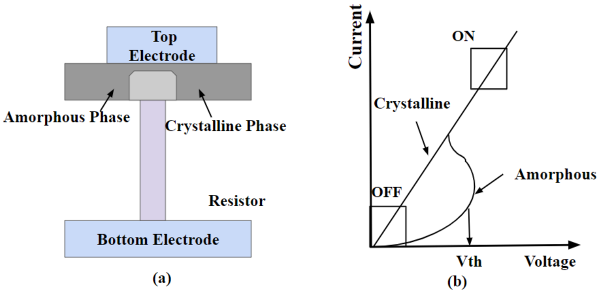

FIGURE 10. PCM: (a) Crystalline state vs amorphous state. (b) I-V characteristics of PCM cells.

The SET and RESET processes are responsible for toggling the device between ON and OFF states. During the transition between ON and OFF states, a brief gap exists in the OFF region owing to the phase-switching process. PCM technology exhibits faster operation, lower power consumption, lower supply voltage, and superior endurance compared to flash memory technology.

C. SPIN-TRANSFER TORQUE MAGNETIC RANDOM-ACCESS MEMORY

Magnetic Tunnel Junction Spin-Transfer Torque (MTJ STT) technology is an important mechanism in Magnetic Random-Access Memory (MRAM), that employs a magnetic tunnel junction that comprising layers of ferromagnetic material separated by an insulating tunnel barrier. In STT MRAM, a magnetic tunnel junction (MTJ) consists of three layers: two ferromagnetic (FM) layers separated by a thin insulating tunneling barrier. Within the MTJ, the fixed layer is magnetically linked with an antiferromagnet (AFM), which requires stable magnetization while undergoing voltage changes. The resistance within the MTJ varies between low when the magnetizations of the FM layers are parallel and high when they are anti-parallel. This variation can be quantified using the tunneling magnetoresistance (TMR) ratio [31].

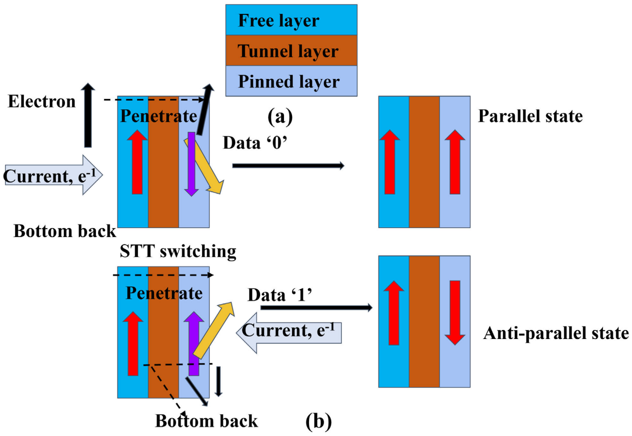

FIGURE 11. Figure caption illustrating the MTJ structure and the process of spin transfer torque-induced magnetization switching: transitioning from an anti-parallel to parallel orientation (top) and from parallel to anti-parallel (bottom).

The spin-transfer torque (STT) influences the magnetization of the MTJ's free layer when an unpolarized electric current becomes spin-polarized upon passing through the fixed layer [32]. Fig. 11 depicts the cell structure of the MTJ STT switching mechanism, transitioning from an anti-parallel orientation to a parallel orientation. During this process, electrons move from the pinned layer to the free layer, where magnetization is easily rotatable. As electrons rotate the magnetization, they align in the same spin direction, generating a spin-polarized current upon traversing the pinned layer. This spin-polarized current then affects the free layer, subjecting it to torque based on spin angular momentum. The magnetic state of the free layer changed when the torque exceed a specific threshold value.

For a transition from a parallel to an anti-parallel orientation (illustrated in Fig. 11(b)), electron flow needs to occur from the free layer to the pinned layer. Electrons that maintain the same spin direction after passing through the free layer and reaching the pinned layer facilitate this

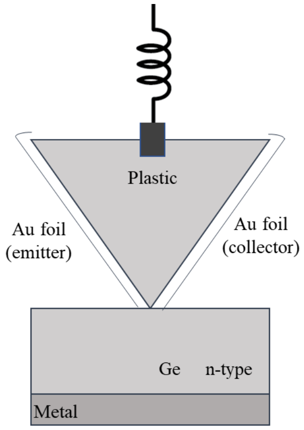

FIGURE 12. Schematic of the point contact transistor.

transition. Conversely, electrons with different spin directions are reflected at the insulator-pinned layer boundary and bounce back to the free layer. This reflection induces a spin transfer torque on the free layer, ultimately switching its magnetization when the torque exceeds the threshold value.

VI. TRANSISTOR TECHNOLOGY IN NEUROMORPHIC DEVICES

A. EVOLUTION OF TRANSISTOR TECHNOLOGY

The first seed of the electronics industry called the 'vacuum tube' was invented in 1904 [33], as a device designed to regulate electron flow within a vacuum. However, during World War II, the demand for vacuum tubes surged, revealing their limitations. These tubes are plagued by increasing complexity, cost, and power consumption which degrade their reliability. As the 1940s drew to a close, the electronic industry discovered two semiconductor devices: the point-contact germanium transistor and the bipolar junction transistor (BJT) [34] as shown in Fig. 12.

In a landmark achievement in 1947, a team comprising William Shockley, John Bardeen, and Walter Brattain introduced the point-contact transistor [35]. A year later, in 1948, William Shockley pioneered BJT. This three-terminal device plays a crucial role in our everyday existence as amplifiers and switch, impacting our lives in numerous subtle yet significant ways. BJTs typically consume more power and have lower switching speeds which makes them less efficient in applications where power efficiency is critical. The transistor's trans-formative legacy extends to the evolution of field effect transistors (FETs). FETs are voltage-controlled devices, that provide a high level of performance in terms of power efficiency and reliability. The FET uses an electric field to control the current flow making it a voltage device instead of a current device.

The very first FET device called the Junction Field Effect Transistor (JFET) patented by Heinrich Welker in 1945, is often used in low to medium-frequency amplification and switching circuits [36]. Despite the invention of the junction field effect transistor, the journey was far from over, as its performance fell short of expectations, leaving room for exploration and improvement in other forms of the device. The JFET device has introduced more gate leakage current owing to high drain voltage which degrades the device's performance. In 1959, Mohamed Atalla and Dawon Kahng [37] discovered the metal-oxide-semiconductor field effect transistor (MOSFET) [38] considered the driving engine of the semiconductor industry. MOSFETs are incredibly versatile 20th-century inventions, which have become iconic for their role in making tiny chips, MOSFETs are used in almost all electronic devices, from amplifiers and voltage regulators to microprocessors and memory cells. They are also essential in power management circuits and switching applications. It packed more into less space, remained affordable, and ran faster. For over 40 years, MOSFETs have dominated the semiconductor industry. Scaling is an important approach, for increasing the packing density of the chip. Scaling enhances the device's speed functionality with a minimum fabrication cost. However, MOSFET faces various unwanted problems owing to reducing the device dimensions. The overscaling of device dimensions in MOSFET introduces a non-ideal effect called short channel effects (SCEs) [39]. These SCEs affect the device's efficiency. In 1997, Dr. Chenming Hu invented a fin-shaped field effect transistor (FinFET) [40]. FinFET is a triple gate device, that has better electrostatic gate control capability over the channel which reduces the leakage current. Intel was the first company to adopt FinFET technology over CMOS technology in 2012. FinFET technology [40] has been used in the semiconductor industry for more than one decade.

FETs are a broad category of transistors that are used in various applications from sensors to memories [41]. The Metal-Semiconductor Field-Effect Transistor (MESFET) [42] is a type of field-effect transistor (FET) commonly used in high-frequency and high-speed applications, particularly in Radio Frequency (RF) and microwave circuits, including satellite communication, radar systems, and wireless communication. MESFETs are typically fabricated using compound semiconductor materials such as gallium arsenide (GaAs) or indium phosphide (InP). These materials offer a high electron mobility, which is advantageous for high-frequency applications.

Graphene FETs (GFETs) [43] offer unparalleled potential as electronic devices in the semiconductor industry, particularly for high-frequency and high-speed applications. Departing from the conventional FET design, GFETs utilize graphene as the channel material instead of silicon, capitalizing on graphene's superior mobility, enhanced thermal conductivity, and lower parasitic capacitance. These unique properties empower the seamless design of cutting-edge RF electronic circuits, making GFETs a promising choice for the future of semiconductor technology. The various FET configurations are presented in Fig. 13.

Transistor technology has undergone a remarkable evolution since the birth of the transistor, propelling the world into

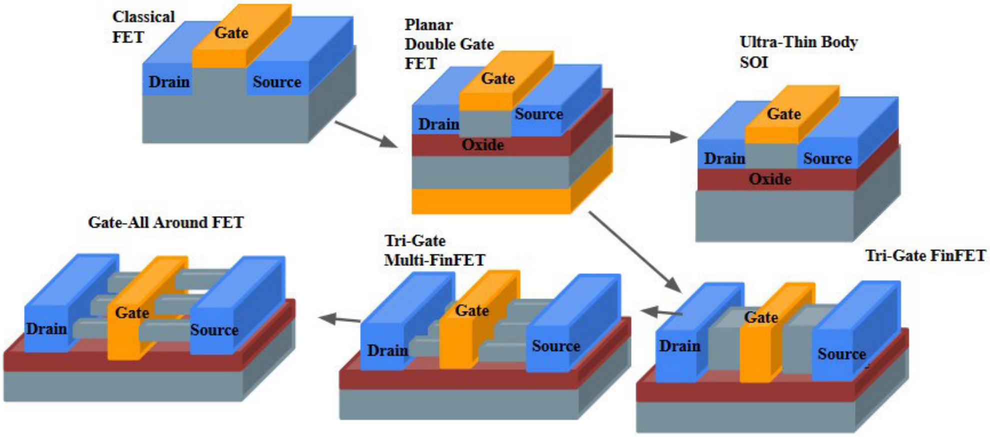

FIGURE 13. Schematic structure of various FETs.

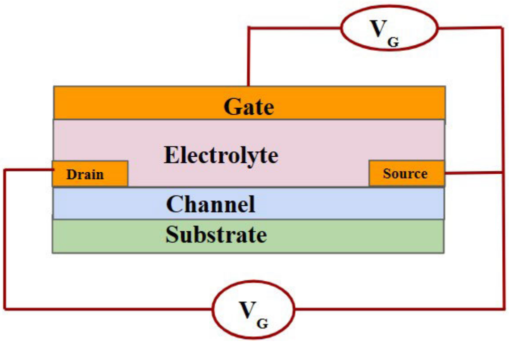

FIGURE 14. IGTs structure.

the digital age. It has come a long way since its inception, transforming the world of electronics and paving the way for countless technological advancements. This demonstrates a remarkable tale of miniaturization, efficiency, and exponential growth. The evolution of transistor technology has revolutionized the computing, communication, and countless industries, leaving an indelible mark in the modern world. Over the decades, transistors have shrunk in size, expanded in functionality, and become an integral part of our daily lives. The evolution of transistor technology from vacuum tubes to silicon transistors and beyond has been a defining factor in the electronics industry. The relentless pursuit of smaller, faster, and more efficient transistors has driven the evolution of technology. Transistors are used in many electronic devices, from tiny parts to big computers. Without them, many of our gadgets would not have worked. In the next section, we provide a detailed overview of the emerging memory devices for a better understanding.

VII. TYPES OF TRANSISTORS USED IN NEUROMORPHIC DEVICES

A. ION-GATE NEUROMORPHIC TRANSISTORS

Ion-gate transistors (IGTs) share a similar structural and voltage bias configuration with MOSFETs as shown in Fig. 14. However, a crucial divergence arises from the choice of the gate dielectric materials. While MOSFETs rely on insulating gate dielectrics, IGTs opt for electrolytes, which serve dual purposes as electron insulators and ion conductors within the gate dielectric. These differences in the gate dielectrics lead to distinct working mechanisms for the two devices. MOSFETs fundamentally control the current flow when a supply voltage is applied to the gate and the gate to the drain terminal. The applied gate voltage determines the extent of control over the channel, allowing MOSFETs to act as crucial switches or amplifiers in electronic circuits [44]. This process exclusively involves electron

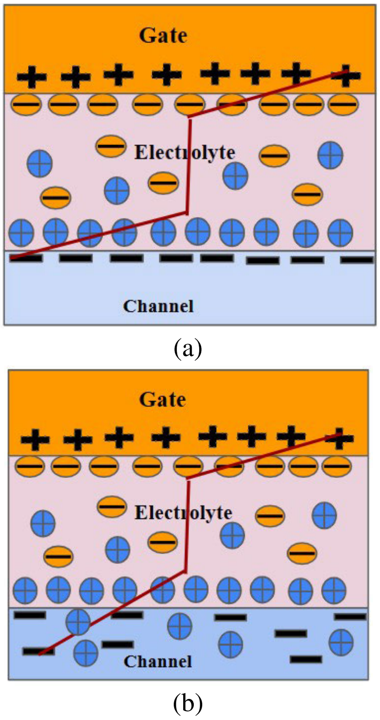

FIGURE 15. (a) Electrostatic doping mechanism with an impermeable channel (b) Electrochemical doping mechanism with a permeable channel.

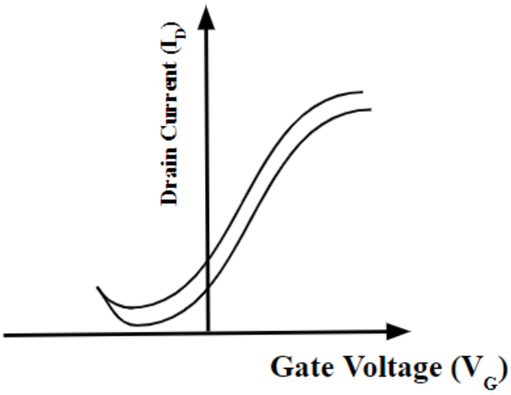

FIGURE 16. Transconductance ($G_m = dI_D/dV_{GS}$) performance of IGTs Structure.

movement and is essentially a capacitive charging mechanism. On the other hand, IGTs follow the same roots as MOSFETs. In IGTs, the gate terminal bias governs the control of the current. However, in IGTs, the movement of ions within the electrolyte of IGTs, come into play at specific gate voltage levels. This ion movement is a critical factor in the distinctive working mechanism of IGTs within the electrolyte, which occurs at specific gate voltage levels. This ion movement is a key factor in the unique working mechanism of IGTs [45].

In IGTs structures, there are two distinct working mechanisms electrostatic and electrochemical carrier doping mechanisms, where impermeable and permeable channel materials can be considered [46]. Fig. 15(a) represents the schematic diagram of the IGTs where an impermeable channel is used. Applying a positive bias to the gate terminal in the presence of an electrolyte results in the attraction of cations to the negatively charged gate, while anions are repelled. While the channel remains impermeable, positive ions from the electrolyte gather near the interface between the electrolyte and the channel. This accumulation of oppositely charged ions creates an electric double layer (EDL), essentially forming a parallel plate capacitor structure, where one plate is formed by the accumulated positive ions and the other plate is the channel surface with induced negative ions. The formation of EDL is a fundamental concept in electrochemistry is widely utilized in various applications, including capacitors, supercapacitors, and electrochemical sensors, in which the charge separation at the interface plays a crucial role.

Fig. 15(b) the working function of the IGT is depicted, taking into account the permeable channel. Upon applying a positive voltage to the gate terminal, electrolyte ions start migrating toward the channel, facilitated by the channel's ability to permit ion passage through the interface. The electrolyte ions are injected more towards the channel owing to the high supply voltages. An opposite type of charge carrier was injected into the channel region to compensate for the injected electrolytic ions. The electrochemical doping mechanism involves the injection of electrolytic ions whereas opposite ions move towards the gate. IGT-based electrochemical doping is also called electrochemical transistor (ECT). IGTs have polarized and nonpolarized gates, where the nonpolarized gate configuration boasts a higher capacitance value that enhances gate controllability compared to a polarized gate setup [47], [48]. The polarized gate can increase the gate controllability of the channel by enlarging the area of the gate-electrolyte interface.

Fig.16 shows the transconductance performance of the IGT transistor, which IGT possesses a significant capacitance value, thereby enhancing the transconductance value in comparison to traditional transistors. This characteristic is advantageous for small signal amplification.

Different types of IGT structures exist in practical applications.They can be categorized into two primary groups: planar and vertical. Planar structures, in turn, can be further classified into co-planar and lateral thin-film transistor (TFT) structures. In the co-plane structure, the source, drain, and gate electrodes are placed in the same plane, as shown in Fig. 17(a). Within co-planar structures, the source, drain, and gate electrodes are commonly deposited simultaneously onto the substrate through methods like spin coating. These IGTs are ideal for sensors because of their simple design, easy fabrication, and the fact that they don't require device miniaturization

Another planar IGT structure lateral thin-film transistor is shown in Fig. 17(b). Lateral TFT-based IGTs have four specific configurations based on the locations of the source, drain, and gate electrodes: top-gate top contact, top-gate bottom contact, bottom-gate top contact, and bottom-gate bottom contact. Unlike co-plane architecture, lateral TFT designs are more mature in their processing and offer integration possibilities into arrays. However, the energy efficiency and processing speed of planar IGT structures need to be improved. To address these issues of planar IGT, a vertical field effect transistor (vFET) [49] has been demonstrated to improve the saturation current and switching frequency, as shown in Fig. 17(c). The vFET structure offers a higher saturation current with minimum operating voltage which reduces the total power consumption of the device [50]. A comparative analysis between the IGT and alternative devices for implementing neuromorphic computing is presented in Table 1.

B. FERROELECTRIC-GATE NEUROMORPHIC TRANSISTORS

Since the discovery of ferroelectricity in $BaTiO_3$, ferroelectric materials, have attracted considerable attention [51]. Ferroelectric field effect transistors (FeFETs) use ferroelectric materials as gate dielectrics. FeFETs typically use the same designs as their conventional counterparts, but instead of relying on an external electric field, they control the conductance of the channel by polarization [52]. The loop in Fig. 18 shows a characteristic hysteresis loop that displays the relationship between the induced polarization and the applied electric field. The hysteresis loop shape

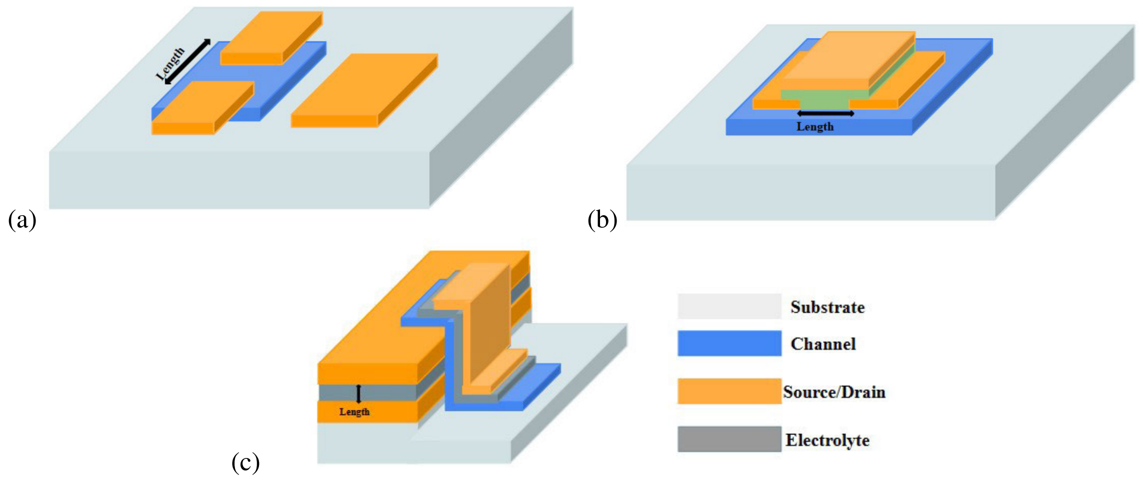

FIGURE 17. The basic structures of various IGTs. (a) Co-plane structure (b) Lateral TFT structure (c) Vertical structure.

signifies the inherent ferroelectric properties of the material including its coercive field (E_c), remnant polarization (P_r), and polarization saturation (P_{max}). The transistor provides features such as a high operating speed and multi-domain switching possibility and hence, is suitable for the application of neural computing. FeFETs can possess two architectures. The Metal-Ferroelectric-Insulator-Semiconductor (MFIS) or Metal-Ferroelectric-Metal-Insulator-Semiconductor(MFMIS). Lue et. al. provided detailed information on the simulation and modelling of these two types of transistors [53]. It discusses the behavior of the drain current, channel potential, surface band bending, and space charge density as functions of the drain voltage. This study uses a wide range of materials and geometric parameters to gain insight into the operation of FeFETs. It also describes the calculation algorithm used and states that the results are independent of the equivalent oxide thickness (EOT) of the insulator. The study mentioned that the hysteresis loop of the ferroelectric layer traces a clockwise direction for a p-type substrate and a counterclockwise direction for an n-type substrate.

The diagram (Fig. 19) illustrates the conventional structure of a FeFET, the gate voltage pulses applied to it, the resulting multi-level polarization states, and the corresponding transfer curves [54]. By applying short voltage pulses to the gate, the threshold voltage of the underlying MOSFET channel can be gradually tuned, thereby adjusting the drain-to-source conductance. This feature allows the implementation of weight updates in FeFET synapses, enabling the storage and manipulation of analog synaptic weights in neural computing systems. The carrier concentration in the channel can be precisely and gradually adjusted by manipulating the polarization state of the ferroelectric dielectric using gate voltage pulses. This ability allows FeMFETs to distinguish various logic states, making them capable of serving as both memory and logic devices with non-volatile characteristics. In the context of neuromorphic transistors, the utilization of multi-domain polarization switching enables multiple levels of channel conductance to be achieved. This capability is valuable for recording the synaptic weights in neuromorphic circuits.

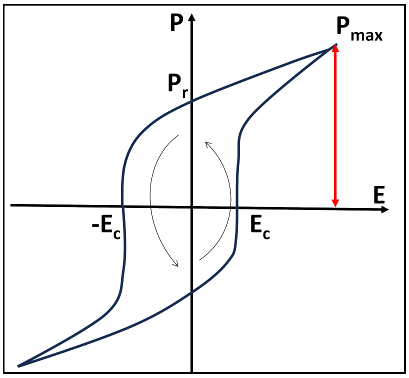

FIGURE 18. Hysteresis loop illustrating the polarization-electric field (P-E) behavior of a ferroelectric material under varying electric fields. The hysteresis loop's shape signifies the material's inherent ferroelectric properties, including its coercive field (E_c), remnant polarization (P_r), and polarization saturation (P_{max}).

The analog conductance modulation behavior in ferroelectric thin-film transistors (FeTFTs), comprising nanoscale ferroelectric materials and oxide semiconductors, was demonstrated by Lee et al. [55]. Precise control of the polarization changes within the nanoscale ferroelectric layer induces conductance modulation and depression characteristics in FeTFTs. These devices exhibit potentiation and depression properties characterized by high linearity, multiple states, and minimal cycle-to-cycle/device-to-device variations. Through simulations employing measured properties, a recognition accuracy of 91.1 % for handwritten digits was achieved by a neuromorphic system featuring FeTFTs as synaptic devices. This study presents a potential avenue for realizing neuromorphic hardware systems employing FeTFTs as synaptic devices. Takagi et. al. observed that, in the

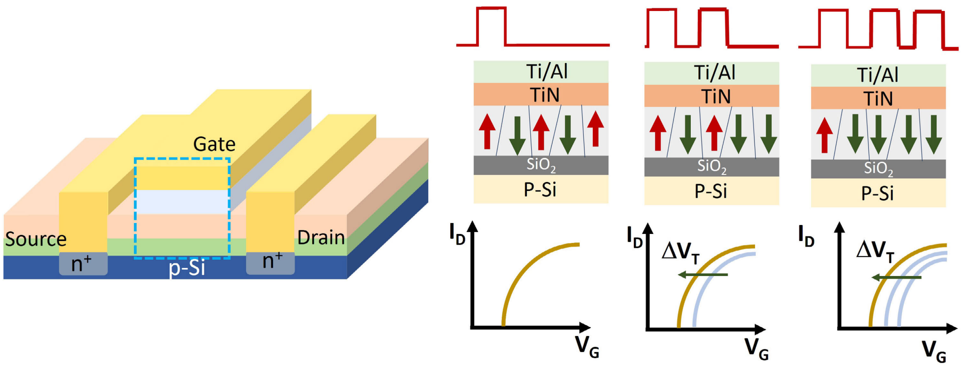

FIGURE 19. The configuration of a FeFET, the applied gate voltage pulses, the numerous switching states of multi-domain partial polarisation, and the associated transfer curves are shown in an illustration [54].

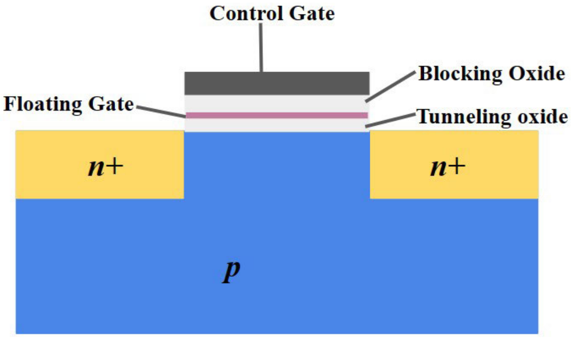

FIGURE 20. Schematic of the structure of FG transistor.

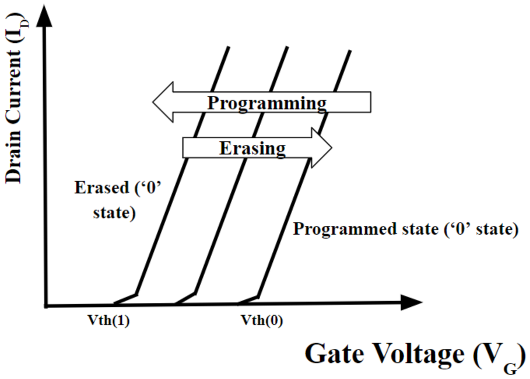

FIGURE 21. I-V characteristics of FG transistor.

case of small polarization regimes, a linear proportionality exists between the memory window (MW) and ferroelectric polarization [56]. The relationship between the MW of ferroelectric field-effect transistors (FeFETs) and the P-E hysteresis loop of a ferroelectric gate insulator was investigated theoretically in this study. Furthermore, it is determined that when the remanent polarization significantly exceeds the product of the permittivity and coercive field, the MW approaches a limiting value equivalent to 2 times the coercive field multiplied by the thickness of the ferroelectric material. Tunneling factors that may influence MW in practical FeFET devices are discussed, including the presence of interlayers, interface charges, and minor-loop operation.

C. FLOATING GATE (FG) NEUROMORPHIC TRANSISTORS

A floating gate transistor is a type of transistor that uses a non-volatile memory device such as flash memory. The first metallic floating gate in a MOSFET was discovered by Kahng and Sze in 1967 [57].

The structure of the FG transistor differs from that of a conventional MOSFET, as shown in Fig.20. A metallic floating gate was placed between two different dielectric layers with two different oxide thicknesses. The thick upper dielectric layer serves as an effective barrier, preventing the flow of charge carriers from the floating gate to the control gate during both programming and erasing processes. On the other hand, a lower thinner dielectric layer can block charge carriers from shifting from the semiconductor layer in the absence of a power supply. Therefore, charge carriers can be stored in the floating gate layer even after a power supply outage [58]. There are many ways in which the charge can shift to or from the floating gate of the transistor such as hot electron injection, Fowler-Nordheim (F-N) tunneling, and direct tunneling.

Fig.21 shows the I-V characteristics of the FG transistor. In an ideal MOSFET, the threshold voltage remains constant at a fixed drain voltage. However, for Floating Gate (FG) transistor devices, the threshold voltage changes because of the trapping of charge carriers in the floating gate. This phenomenon affects the conductivity during programming and erasing operations. When the threshold voltage value is designated as'0' or $V_{Th}(0)$, the Floating Gate (FG) is recognized as the 'programmed state' resulting from the injection of a negative charge. On the other hand, 'erased state' can be called when threshold voltage returns to V_{Th} (1) due to the supply of negative voltage at the gate electrode. In the erased state, no charges were trapped in the floating gate, signifying the 'OFF' or '0' state, while the programmed state is referred to as '1' or 'ON' state.

During hot electron injection programming, vertical and lateral electric fields must be applied. The lateral electric

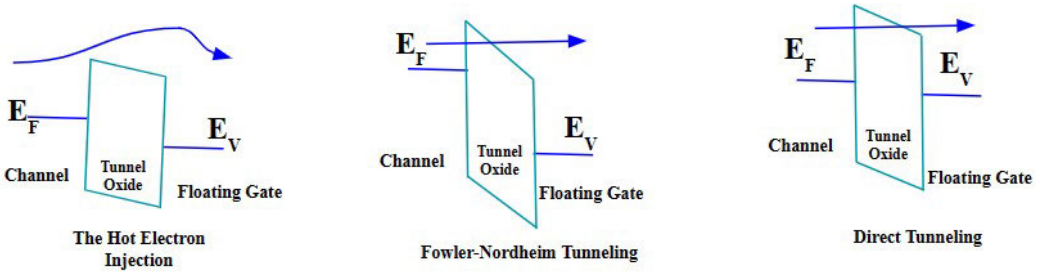

FIGURE 22. Schematic structure of hot electron injection, Fowler-Nordheim tunneling, direct tunneling.

field provides sufficient energy to surpass the energy barrier separating the floating and semiconductor layers. The vertical electric field is generated by the gate voltage, which helps the charge carriers to be trapped in the floating gate layer. This process enables exceptionally rapid write speeds, typically within the microsecond range for a single data bit. Additionally, it enables write operations with significantly lower control gate voltages, making the memory efficient and responsive [59]. The Fowler-Nordheim (F-N) tunneling mechanism occurs because of the presence of a higher electric field, and charge carriers pass through the thin barrier. The thin barrier allows charge carriers to tunnel through it [60]. The F–N tunneling mechanism requires minimum energy for the process of 'program' and 'erase' operation and offers good efficiency compared to other injection methods. However, this process has some disadvantages owing to the high electric field and long access time. The direct tunneling (DT) mechanism occurs because of a higher electric field and extremely thin layers. Owing to the ultra-thin nature of the oxide layer in the DT process, the charge carriers can easily move. The direct tunneling method provides a faster programming speed with less power consumption. However, it reduces the data retention capabilities. To enhance charge retention, increasing the barrier height can be beneficial, which in turn reduces tunneling probabilities. In the DT mechanism, the tunnel oxide layer is generally constrained to around 6 nm. However, because of trap-assisted electron tunneling induced by oxide aging, a more realistic thickness may need to be increased to approximately 7-8 nm [61]. The schematic structure of the hot electron injection, Fowler-Nordheim tunneling, and direct tunneling mechanism ha shown in Fig.22. Conventional FG transistor faces several challenges owing to the over-scaling of the device dimensions. One main problem is the decreasing distance between cells, leading to cell-to-cell interference and parasitic capacitance. The reduction in oxide thickness enhances the tunneling issue which increases the leakage current, making it difficult to manage charge retention. Additionally, a smaller floating gate carries a small number of free electrons which degrades the performance of the device. Scientists have created several enhanced iterations of FG transistors aimed at improving their efficiency. Examples include silicon-oxide-nitride-oxide-silicon (SONOS) [62], and nano-floating-gate (NFG) memory devices, which incorporate metal nanoparticles (NPs) and utilize organic/inorganic nano-materials in their dielectric layers [63]. All the modified structures of the FG transistor offer better durability, and less power consumption with a smaller chip size than conventional FG transistor. Modified structures can store multiple levels of data which attract the semiconductor memory industry. Liu et al. [64] constructed an organic FET (OFET) memory incorporating self-assembled gold nanoparticles (NPs) into the gate dielectric. Constructed on a silicon (Si) substrate to control the gate electrodes, the device utilized a 100 nm thick silicon oxide (SiO_2) layer as the charge-blocking dielectric. Poly(3-hexylthiophene) (P3HT) serves as a semiconductor channel layer, with additional polyelectrolytes separating it and a poly(4-vinylphenol) (PVP) tunneling layer covering the gold nanoparticles (NPs). Despite exhibiting an impressive switching ratio of 1500, the device exhibits a relatively short retention time of 200 s. Ryu et al. [65] approached a non-volatile transistor memory in which a double-stacked layer of metal NPs was used, as shown in Fig.23. They formed various types of charge-tapping layers by depositing different sequences of gold (Au) and nickel (Ni) such as Ni/Ni, Au/Au, Ni/Au, and Au/Ni to observe the memory device performance. The utilization of both top and bottom charge-trapping layers (Au/Ni) has been observed to improve program/erase speeds significantly and notably extend data retention times. Chang et al. [66] demonstrated an FG memory device incorporating high-k oxide dielectrics, such as HfLaO, HfON, and HfO as the blocking, charge trapping, and tunneling dielectric layers, respectively. Memory devices provide a low program/erase voltage of 12 V and also contribute a programming speed of 1/100 ms.

Loai et al. introduced an energy-efficient memristive device-based FG FET device [38]. Operating in a sub-threshold memristive mode, this distinctive device is referred to as the Y-flash [8], [34] and is engineered to be linearized for small change in signals. By incorporating recent advancements in memristive techniques utilized in small-scale selector-free dense integrated Artificial Neural Networks (ANNs) for spike-timing-dependent plasticity (STDP), Vector-Matrix Multiplication (VMM), associative memory, and classification training, we developed a practical and high-performance memristive device.

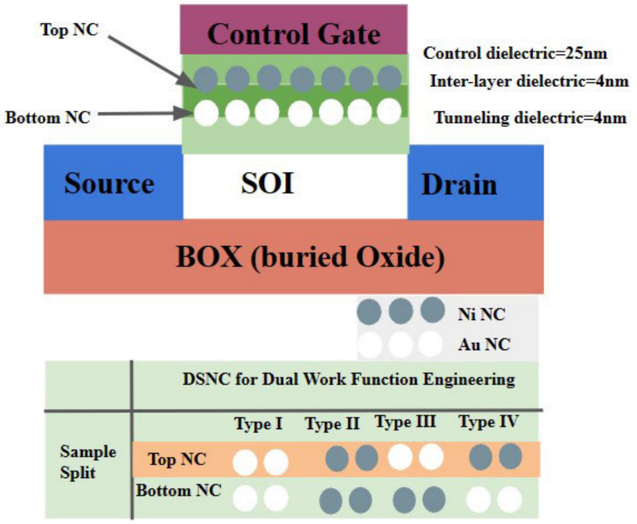

FIGURE 23. Schematic diagram of double-stacked metal nanocrystals (NC).

Through both theoretical analysis and experimental validation, our research demonstrated the viability of this memristive device for applications in high-performance neuromorphic computing. This innovative approach not only paves the way for energy-efficient computing but also opens up new possibilities for advanced neural network applications, making it a promising technology for the future of artificial intelligence and cognitive computing.

D. PERSPECTIVE ON TRANSISTORS IN COMPUTING

Integrating various types of transistors into neuromorphic computing systems offers a holistic approach to emulate the complexity of the human brain. By combining their unique characteristics such as ion-based modulation, non-volatility, and analog behavior—neuromorphic systems can achieve energy-efficient, adaptive, and self-learning capabilities. These advancements have brought us closer to realizing artificial intelligence models that can learn, adapt, and process information in ways that resemble human cognition.

VIII. INTEGRATION OF TRANSISTORS IN NEUROMORPHIC SYSTEMS

IGTs have evolved from traditional MOSFET, that were initially designed as switches for both analog and digital circuits. However, IGT can be used for sensor and neuromorphic computing. IGTs have a large channel capacitance which enhances the transconductance value, and can be applied for small signals.

IGT-based electrostatic and electrochemical doping mechanisms can make neural networks to study synapses in the brain. Lenz et al. [85] demonstrated a vertical structure IGT. Initially, they designed an Au/Ti/SiO_2/Ti/Au stack with a cross structure and created empty spaces perpendicular to the stacking orientation with Ti and SiO_2 materials. The height of these voids corresponds to the length of the channel and their depths, determined by SiO_2 thickness and etching time, respectively, were precisely controlled. The proposed device uses a spin coating technique to fill the constructed voids. It is observed that the proposed vertical IGT improved the high on-state current densities and switching ratio by <3mA cm^{-2} and 10^8, respectively. The power consumption has been optimized by 100 fJ per event which is good compared to other similar types of devices.

TABLE 1. Comparative analysis between IGT and alternative devices for implementing neuromorphic computing.

Device name	Maximum no of states	Maximum operation speed	Maximum operating energy	Non-linearity	Maximum endurance	Maximum retention
IGT	10000 [67]	5ns [68]	1.23fJ [69]	Low	10^9 [70]	>25 h [71]
PCM	26 [72]	700ps [73]	1 pJ [74]	High	10^{11} [75]	>1000 years @ RT [76]
MOSFET	2 [3]	$<10^1$ns DRAM [3]	1–10 fJ [3]	High	$>10^{16}$ [3]	1–10 ns [3]
RRAM	64 [77]	85ps [78]	115 fJ [79]	High	10^{12} [80]	>1000 years @ RT [81]
FTJ /FeFET	10 [82]	10ns [82]	100 fJ [82]	High	4x10^6 [83]	3 days [84]

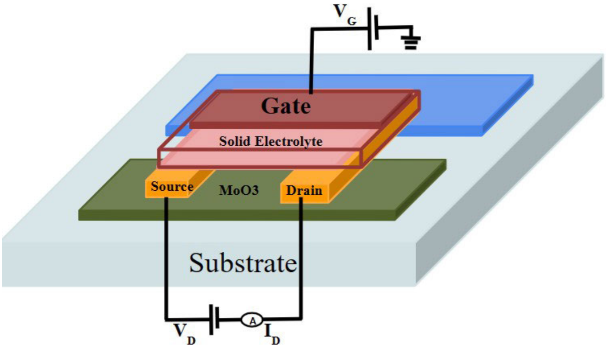

FIGURE 24. Schematic of the structure of IGT based on MoO_3 material.

Shang et al. [86] designed an IGT based on layered transition metal oxide -phase molybdenum oxide (α-MoO_3) as shown in Fig. 24. The higher band-gap characteristics of α-MoO_3 make it an insulator. They considered Li^+ electrolyte solid material instead of a liquid material. When a supply voltage was applied to the gate terminal, tiny particles called Li^+ to move in and out of the α-MoO_3 material. The proposed transistor with Li^+ as the dopants offers a channel conductance switching value of $< 10^{-5}$ Torr. These specialized devices are highly energy-efficient and can be employed to construct large-scale computer systems by densely packing them together.

Fig. 25 schematically illustrates a crossbar array comprising a synaptic-weight layer. The weighted summation of the neural network is emulated in the arrangement by calculating vector-matrix multiplication. The switching of

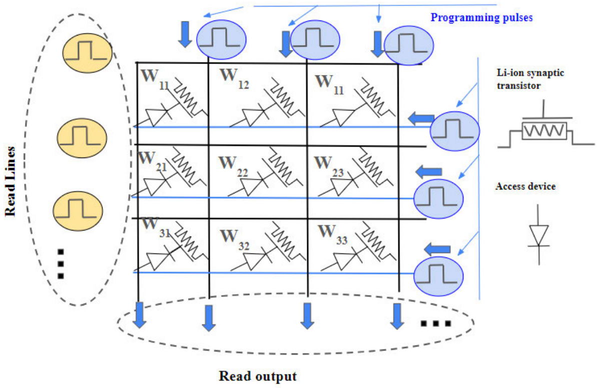

FIGURE 25. Diagram of the synaptic weight layer comprising an IGT crossbar array.

the devices can be performed by programming by applying suitable voltage pulses along the horizontal and vertical lines. Synaptic transistors, based on α-MoO_3, serve as memory elements within the crossbar array [46]. By applying a gate voltage, the interaction of lithium ions with the layered α-MoO_3 channel results in analog switching of the α-MoO_3 layer. The ionic liquid was subsequently substituted with a solid-state Li^+ electrolyte. Li^+ electrolytes are considered dopants that enhance the switching of the channel conductance under vacuum conditions. Additionally, the α-MoO_3-based synaptic transistors demonstrate exceptionally low conductance (<75 nS), making them highly advantageous for energy efficiency and the manufacture of extensive crossbar arrays.

Zhu et al. [87] proposed a latterly coupled IGT featuring a co-planar architecture that operates on an electrostatic doping mechanism. This structure changes the electrical properties to mimic how synapses in the brain work for short-term memory. The planar structure of an IGT can be built with multiple gates and channels. Recently, researchers have been paying considerable attention to the electrochemical doping mechanism of IGT owing to its non-volatile characteristics. This method does not forget information when the power is turned off, which makes it useful for mimicking the long-term memory of the synapses in the brain. Burgt et al. [88] demonstrated a neuromorphic organic transistor based on an electrostatic doping mechanism with affordable plastic materials that act like an artificial synapse [89]. The proposed transistor provides non-volatile and reproducible states (more than 500) at a very low operating voltage, as noted in [90]. However, the devices exhibited a higher channel conductance value ranging from 500 to - 2000 μS.

More electrical power was required to carry the current capacity when the array dimensions were increased. Nevertheless, by combining a conductive polymer with an insulator, the device consumes less electricity. This mixing composition offers a lower synaptic weight readout (< 10 nA) [91].

Wang et al. proposed a ferroelectric material-based transistor for applications in neural computing [92]. The device structure of ferroelectric channel transistors (FeCTs) is shown in Fig. 26. In the device fabrication process, the bottom dielectric layer, aluminum oxide (Al_2O_3) is deposited onto the substrate using atomic layer deposition (ALD). Mechanical exfoliation was used to create a 2D - In_2Se_3 channel layer and bottom hexagonal boron nitride (h-BN). The elimination of the PVA sacrificial layer employed wet methods, while the design of the electrode was accomplished through the utilization of electron beam lithography. Al_2O_3 layers, produced by ALD, dopes the In_2Se_3 channel. Therefore, the bottom h-BN layer's existence is essential for optimizing the interface. In addition, the h-BN layer serves as a passivation layer for the In_2Se_3 channel in addition to being a dielectric, thereby isolating it from the ambient environment. When a voltage is applied to the gate electrode, the polarization of the ferroelectric material changes, resulting in a change in the threshold voltage of the device. This change in threshold voltage allows for the storage and retrieval of information in FeFETs. The ferroelectric material acts as a non-volatile memory element, enabling the FeFET to retain its state even when power is removed. The device's impressive performance includes a 40 nanosecond write speed, increased endurance made possible by the internal electric field, low energy consumption at 234/40 femto-joules per event for excitation/inhibition, and a high-precision simulation with an accuracy rate of 94.74%.

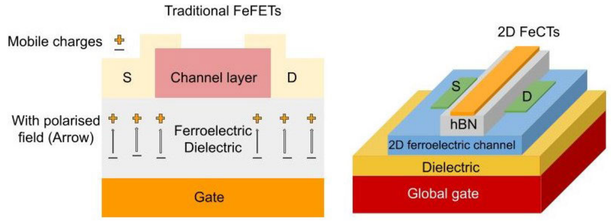

FIGURE 26. Schematic diagram representing the Ferroelectric field effect transistor structure [92].

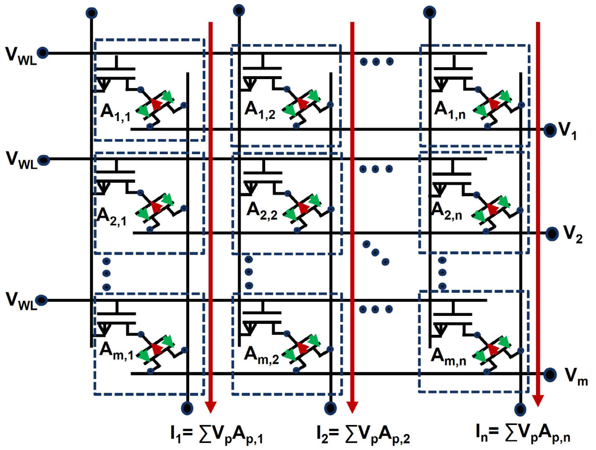

FIGURE 27. Implementation of analog vector-matrix multiplication and row-wise parallel weight updates via a Ferroelectric Field-Effect Transistor (FeFET) pseudo-crossbar array. Incorporating access transistors alongside FeFET storage devices in synaptic weight cells to minimize disturbance effects [54].

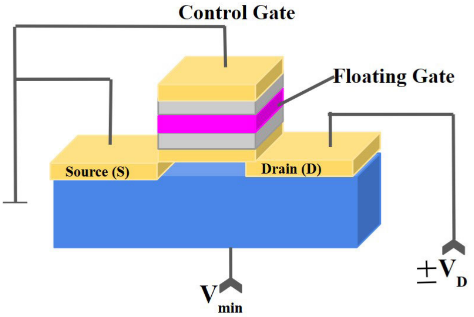

FIGURE 28. Schematic diagram of FG transistor as two terminal devices.

The FeFET synapse can be incorporated within pseudo-crossbar arrays, enabling simultaneous row-based adjustments of the weight and column-based computation of the combined FeFET conductances. Fig 27 shows the crossbar array for the ferroelectric material-based transistors. Using high-speed weight update pulses, which are mapped to the conductance states that vary between dynamic changes, The FeFET synapse combines the ability to modulate the ferroelectric polarization with the help of metal oxide semiconductor field effect transistors. The crossbar arrangement of the FeFET synapses allows the weighted summation of individual FeFET conductance values. The conductance states were programmed, and the results from the columns were read using programming voltage pulses and the application of read voltages.

In the area of neural computing 2D materials are used as both switch and memory components of the neural network architecture [93]. Different 2D materials such as Graphene, MoS_2, WSe_2 etc. been proposed, which can be used in FeFET fabrication to enhance the properties [94], [95], [96]. These materials have been used as the channel materials, gate contacts, and gate dielectrics in FeFETs. These materials contribute to improved device performance, including increased retention time, memory window, and on/off ratio [97]. The use of 2D materials in FeFETs helps address issues such as gate current leakage and charge trapping at the semiconductor-ferroelectric interface. The van der Waals structures of 2D materials effectively reduce or eliminate charge trapping, leading to an enhanced device performance. Wan et al. studied the effect of vertically stacked graphene with hexagonal boron nitride and $-In_2Se_3$, to form a device structure. The electric polarization of $-In_2Se_3$ induces doping in graphene, leading to modulation of its resistance and allowing for the storage and retrieval of information in the FeFET [98]. Si et al. reported the stable non-volatile memory property of these Fe-FETs, which is modulated by the back-gate bias of the MoS_2 transistors, leading to an enhancement of the on/off current ratio. Additionally, the $CuInP_2S_6$ thin film used in these Fe-FETs also shows resistive switching characteristics with a high on/off ratio between the low- and high-resistance states [99]. Tathagata et al. [100] proposed and fabricated a charge tunneling-based synaptic transistor where two-dimensional molybdenum disulfide (MoS_2) is considered as the channel material. MoS_2 exhibits better coupling with metallic gates. MoS_2 can be used for non-volatile memory cells owing its high switching ratio and low band gap characteristics. The proposed architecture offers a high drain current with a sub-threshold swing of 77mV/decade. In 2014, Riggert et al. [101] designed an FG transistor as a memristive device for neuromorphic computing as shown in Fig.28. Their investigation focused on evaluating the impact of gate oxide scaling in single floating-gate transistors on their performance in the memristive operation mode within the domain of neuromorphic engineering. It is considered that an oxide thickness of 4nm can generate a pulse width of 3ms, which is comparable to the biological pulse times that consume less power than other memristive devices. In investigating thinner gate oxides, they formulated an enhanced device model for a single MemFlash cell, encompassing F-N tunneling, Poole-Frenkel emission, and hot electron injection. The DT tunneling process has been used to consider an ultra-thin oxide layer from the channel to the floating gate.

Yongli et al. [102] demonstrated indium-gallium-zinc-oxide (IGZO) based floating-gate synaptic transistors (FGST) for neuromorphic computing where the minimum temperature is considered, as shown in Fig.29. The proposed transistors used Al_2O_3/ITO/Al_2O_3 gate dielectric stacks to store the synaptic weight as channel conductance (G). They designed an artificial neural network with 95.7% accuracy for small signal datasets. Fig. 30 illustrates a cross-bar array composed of FGST. They examined M × N synaptic devices designed to mimic an M × N synaptic weight array. The conductance state of these synaptic devices is determined by the product of high voltage with the rows and columns lines, with the rows represented by blue horizontal lines and the columns by black vertical lines. However, reading operations can account for lower voltages with read lines. The crossbar array of AXB synaptic devices emulates the AXB synaptic weight array. The conductance states of the FGST devices were programmed using the voltages along the horizontal programming lines and vertical lines to perform vector-matrix multiplication. By applying small voltages across the read lines (indicated by black horizontal lines), reading operations were performed, and a small voltage was applied across the programming lines to measure the currents from the columns. To prevent discharging after programming, a two-terminal access device is essential for the floating gate transistor device.

Myung et al. [103] proposed a modified FG transistor called overturned charge injection synaptic transistor (OCIST) specifically for neuromorphic computing. The key distinction of OCIST lies in the incorporation of an additional layer known as the Charge Valve Layer (CVL), setting it apart from the conventional FG transistors. CVL plays a pivotal role in regulating the flow of charge carriers to the Floating Gate (FG). The experimental results demonstrated the effectiveness of this modification, with the proposed OCIST achieving an impressive accuracy rate of 92.4% for handwritten digits in the MNIST dataset. This high accuracy underscores the practical viability and potential applications of the OCIST in the field of neuromorphic computing. Comparative analysis of the performance of various neuromorphic transistors has tabulated in Table 2.

FinFETs and other advanced devices have gained prominence in contemporary neural computing. Dibyendu et al. [17] introduced a novel bulk FinFET design centered on ultra-low energy artificial neurons, demonstrating a comprehensive comparative analysis with other CMOS-compatible devices. Avinash et al. [104] designed an energy-efficient bipolar I-MOS for spiking neural networks which reduced the spike energy and enhanced the spike frequency by an order of 6 compared to biological neurons. Neha et al. proposed a junctionless FET configuration-based leaky integrate-and-fire (LIF) neuron. It is observed that the proposed device with a gate length of 20nm consumes less spike energy on the order of 1.14 pJ making the device more power efficient than the partially depleted (PD) SOI MOSFET device.

IX. EMERGING MATERIALS IN TRANSISTOR FABRICATION

A. SI BASED TRANSISTORS

The most important building block used in the production of transistors is silicon wafers. The semiconducting property of Si and its tunability with doping facilitate its application in transistors, which are crucial components of electronic devices. Single crystalline Si wafers are ideal for high-performance transistors owing to their well-ordered atomic structures. Due to the superior thermal conductivity of silicon, the heat produced during transistor operation can be effectively dissipated [111]. Si is suitable for a variety of fabrication processes, including oxidation and etching, because they are chemically stable and do not react with the majority of common chemicals. Because Si wafers are mechanically strong, they can endure the strain of different production processes including lithography and etching [112], [113]. The preparation of silicon wafers is the first step in the manufacturing of transistors. The requisite layers and structures for transistors are created on wafers by several procedures, including oxidation and epitaxial growth after they have been cleaned to eliminate impurities. A sophisticated circuitry of transistors and other semiconductor devices is created on top of these silicon wafers.

Wagner et al. discussed the use of silicon for thin-film transistors (TFTs) in the industrialization of flexible backplanes [114]. The two main research directions for TFTs are processability on flexible substrates and sufficient field-effect mobility of electrons and holes. Different modifications of silicon films, such as amorphous, nanocrystalline, and microcrystalline, are summarized in terms of their TFT properties and compatibility with foil substrate materials.

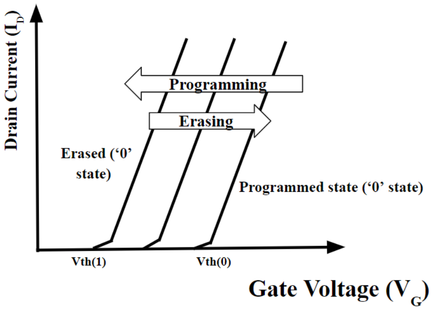

FIGURE 29. A schematic diagram of IGZO-based FGST.

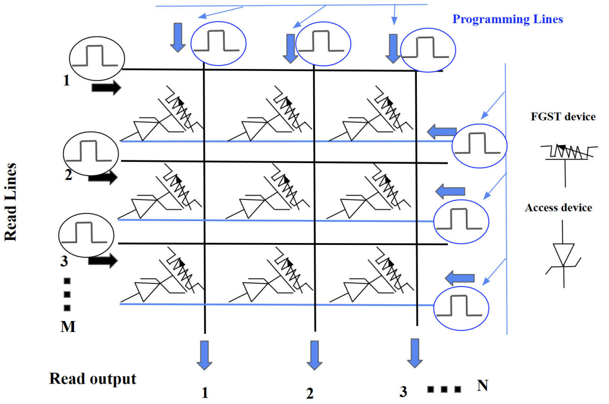

FIGURE 30. Hardware implementation of cross-bar array comprised of FGST.

TABLE 2. Comparative analysis on the performance of various neuromorphic transistors. CNT:Carbon NanoTube, IZO: Indium Zinc Oxide, PEG: Polyethylene Glycol, ITO: Indium Tin Oxide, NA: Not Applicable.

Transistor type	Gate dielectric layers	Channel Material	Spike voltage	Spike duration	Power Consumption	Ref.
IGT	PEDOT:PPS/ Nafion	PEDOT: PSS/PEI	1.5V	1s	10pJ	[71]
	Nano-granular SiO_2	IZO	0.3V	10ms	45pJ	[87]
	PEG	CNT	5V	1ms	7.5pJ	[105]
	Chitosan	IGZO	2V	25ms	1nJ	[106]
Fe FET	$HfZrO_x$	IGZO	4.3V/-3.6V	10ms	NA	[55]
	P(VDF-TrFE)	Pentacene	10 V	500 ms	37.95 nJ	[107]
	$HfZrO_x$	IGZO	-8V/+8V	$3\mu S/3\mu S$	NA	[108]
FGT	Al_2O_3/ITO/ Al_2O_3	IGZO	-4V/+4V	25mS	0.4pJ	[102]
	SiO_2/ graphene/ hBN	MoS_2	-7V/+7V	100 ms	NA	[109]
	SiO_x	CNT	-8V/+8V	100 ms	NA	[110]

This document presents a chart showing the compatibility of directly deposited silicon channels with different substrate materials as shown in Fig. 31. Silicon films, including amorphous, nanocrystalline, and microcrystalline, are leading

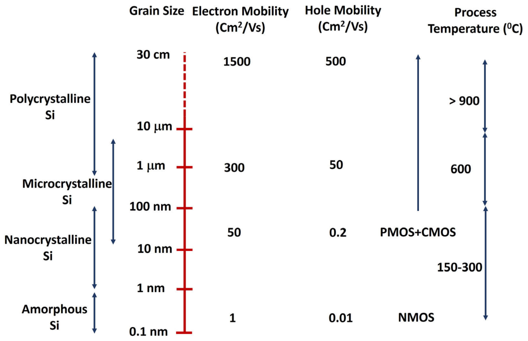

FIGURE 31. **Variations in silicon film modifications accessible for Thin-Film Transistor (TFT) fabrication exhibit diverse carrier mobilities, Field-Effect Transistor (FET) capabilities, and process temperature requirements [114].**

candidates for flexible and conformal TFT backplanes. The choice of silicon material, substrate material, and fabrication process will play a crucial role in the successful implementation of TFT backplanes for various applications. Guo et al. reported the fabrication of FET based on Si using nanoimprint lithography [115]. With the help of lithography techniques and dry etching, they have patterned transistors in silicon including nanowire channels, quantum dots, and ring structures without any degradation in the device structure and characteristics. This method of nanoimprint lithography thus helps in the mass production of transistors for industrial-scale applications.

Ast et al. conducted a comprehensive study to explore the conduction mechanism and sources of leakage current in undoped channel polycrystalline silicon thin-film transistors (TFTs) produced under diverse processing conditions. Remarkably, they achieved leakage currents below 1 nA at drain-source voltages of 40 V for both n-type and p-type devices. The effective channel mobilities for electrons and holes were determined to be 75 and 42 cm^2/Vs, respectively.

B. FERROELECTRIC MATERIALS BASED TRANSISTORS

Ferroelectric materials have a special dielectric characteristic because they retain a constant polarization [116]. These substances can change the direction of their dipoles when exposed to an external electric field [117]. The perovskite structure (ABO_3), in which A and B are cations and A, B, and oxygen atoms are located at the corners of the crystal, body center, and face centers, respectively, is the structure most commonly used for ferroelectric materials [118], [119]. Along with their ferroelectric properties, these materials exhibit good magnetic and optical properties [120], [121]. The B atom moves when an external electric field is applied, causing an unbalanced distribution of electrical charges and formation of a dipole moment. Depending on their symmetry properties, the materials were classified into 32 different crystal classes. 21 out of the 32 crystal classes found in the materials are non-centrosymmetric. Twenty of these non-centrosymmetric crystal classes exhibit piezoelectricity,

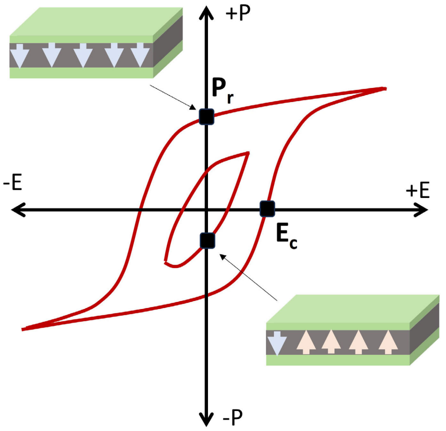

FIGURE 32. **P-E loop characteristics corresponding to different switching states [92].**

while 10 exhibit pyroelectricity, or temperature-dependent polarization [122]. The polarization vs electric-field graph of some pyroelectric materials also exhibits external electric-field-dependent polarization, which results in a hysteresis loop. Crucial characteristics of ferroelectric materials are obtained from this hysteresis loop as shown in Fig. 32 [123]. Ferroelectric materials must exhibit saturation polarization (Ps), remnant polarization (Pr), and coercive electric field (Ec). The values Ps and Pr represent the maximum polarization attained in the presence of an external electric field and the remaining polarization after the field has been removed. Where Ec represents the strength of the electric field required to cause a polarization reversal. Polarisation reversal in ferroelectric materials occurs because of the formation and growth of ferroelectric domains. Ferroelectric memory is an important requirement for non-volatile memory with stable memory states because the polarization switching process is controlled by the electric field and maintains its state even after the field is withdrawn [124].

Ferroelectric materials exhibit significant promise as potential candidates for synaptic weight elements in neural network hardware owing to their non-volatile multi-level memory effect. Ferroelectric materials have characteristic features such as low symmetry and the presence of spontaneous polarization states [125], [126], [127]. Materials such as HfO_2, $PbZrO_x$ (PZT), $SrBi_2Ta_2O_9$ (SBT), $BaTiO_3$ (BTO), and $BiFeO_3$ (BFO) can be used for ferroelectric transistor fabrication owing to their potential for neuromorphic applications. This characteristic holds vital significance for their application in mobile devices, where vector-matrix multiplication is performed during portable artificial intelligence services. Additionally, the adaptive learning effect observed in ferroelectric polarization has attracted substantial research attention, with the aim of improving the CMOS circuit overhead associated with integrators and amplifiers featuring activation functions. Material-related challenges have been identified as potential hurdles in the commercialization of these devices, particularly in CMOS processing and device

TABLE 3. Material parameters used for transistor fabrication.

Category	Materials	Band gap (eV)	Electron Affinity (eV)	Dielectric constant	Device
Amorphous	SiO_2	9 [130]	0.85 [131]	3.9 [132]	Ion Gate Transistors, FeFET, Floating gate transistors
	AlO_x	6.7 [133]	3.7 [133]	9 [133]	
Ferro electric materials	HfO_2	5.8 [131]	2 [?]	23 [134]	FeFET
	In_2Se_3	1.4 [135]	3.6 [135]	17 [136]	
2D materials	hBN	5.95 [137]	1.14 [138]	5.06 [139]	Ion Gate Transistors, FeFET, Floating gate transistors
	MoS_2	1.3 - 2* [133]	4.7 [140]	3.93-10.5 [133]	

* Varies depending on the number of layers

structures [128]. Ferroelectric materials exhibit properties such as high operating speed, low power consumption, and nondestructive read-out for transistor fabrication which is suitable for neural computing applications. Xue et al. reported the multi-domain polarization switching behavior of ferroelectric materials, which helped achieve multi-level FeFET channel conductance [129].

Seidel *et.al.* presented the advantages of utilizing hafnium oxide-based FeFETs for non-volatile memory device applications [141]. Their fundamental three-terminal structure enables the selective activation or deactivation of specific devices and allows for the tuning of the linearity and dynamic range to cater to specific applications. Furthermore, the article delves into the influence of material properties on the ferroelectric layer, interface layer thickness, and scaling on device performance. Notably, it demonstrates the viability of achieving favorable device properties even in the case of highly scaled devices, as small as 100 nm x 100 nm. Another material, In_2Se_3 has been explored as a ferroelectric semiconductor channel material for transistors. It demonstrates strong ferroelectric properties at ambient room temperature and can maintain ferroelectric polarization even at the atomic scale. This substance provides compact, scalable devices that combine non-volatile memory (NVM) with neural computing capabilities. Table 3 lists the different materials and their properties that must be considered for transistor applications.

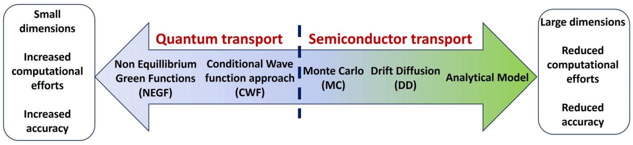

FIGURE 33. Hierarchical ranking of electron device modelling approaches based on computational demand and accuracy [143].

C. 2D MATERIALS BASED TRANSISTORS

Two-dimensional (2D) materials, such as hBN and transition metal dichalcogenides (TMDCs), play a crucial role in the development of transistors for silicon technology. They offer unique properties and have potential applications in next-generation computing technologies. Utilizing 2D materials in transistors presents numerous benefits, such as substantial memory hysteresis windows, prolonged retention, increased endurance, rapid write speed, flexible neuroplasticity adjustment, and exceptionally low power consumption. Additionally, they demonstrate thermal tunability in both memory and neural computation, positioning them as promising candidates for high-density, energy-efficient memory and computing fusion systems. The incorporation of 2D materials in transistors opens avenues for the advancement of high-density and energy-efficient memory computing systems, offering potential solutions to eliminate the physical separation of memory and computing. This addresses challenges in data-centric applications and enhances overall system efficiency.

hBN and TMDCs have attracted significant attention because of their exceptional electrical and mechanical properties. They can be used as channel materials in transistors, thus enabling high-performance and low-power devices. The device physics of ferroelectric transistors based on graphene and TMDCs involves understanding the complex interface interactions between the ferroelectric gate and nanoscale channel. For graphene-based transistors, the focus is on interfacial screening dynamics, and mobility limits at different temperatures. The key benefit of 2D materials is their thermodynamically stable nature as single atomic layers. These materials can be fabricated by isolating them from closely related 3D variants, that have a layered structure. Single 2D layers do not have any dangling bonds and form defect-free van der Waals interfaces with other layered materials. This allows for high charge carrier mobility and minimizes scattering at the surface traps. The thinness of 2D materials, with a thickness below 1 nm, suppresses short channel effects and enables the scaling of the channel length below 15 nm, which is required for technology nodes beyond 3 nm technology nodes. Additionally, the van der Waals interface between adjacent layers creates the possibility of combining different 2D materials in van der Waals heterostructures, offering a wide variety of options for creating novel device designs [142].

Different models have been proposed to provide insights into the performance, reliability, and stability of FETs based on 2D materials [Fig. 33]. Non-equilibrium green's function

(NEGF) is a quantum mechanical approach used to describe the behavior of charge carriers and energy exchange in transistors [12]. It is based on the open boundary Schrödinger equation and is commonly used to model devices in the ballistic regime. The NEGF requires the calculation of the Hamiltonian, Green's function, and self-energy matrix. It is typically used to study nanoscale electronic devices and has been applied to simulate transport in various materials, including graphene, transition metal dichalcogenides (TMDs), and semiconducting heterostructures. The drift-diffusion model is a simplified approximation of the Boltzmann transport equation (BTE) and is computationally efficient. The model is based on the balance equation of the charge carrier flow within the phase space and is formulated as a partial differential equation for the carrier distribution function. These relations consider the effects of mobility, diffusion, and potential distribution in the semiconductor. The drift-diffusion model is commonly used to simulate current transport in prototype FETs based on 2D semiconductors with micrometer scale dimensions [142].

X. ADVANTAGES OF EMERGING TRANSISTORS FOR NEUROMORPHIC COMPUTING

IGTs exhibit elevated transconductance, rapid speed, and the ability to be independently gated, allowing for the creation of a scalable and adaptable integrated circuit. [144]. IGTS can operate at low voltage (less than 4V) owing to the making of a dual layer at the junction between the oxide semiconductor and the ion-gating medium and provides a high charge carrier density due to its high capacitance value [145]. IGTs exhibit significant channel capacitance because of the ion movement mechanism, resulting in substantial transconductance when compared to traditional transistors. This characteristic makes them well-suited for amplifying small signals and for processing neural signals, especially in scenarios where weak signals need to be accurately detected and amplified. IGTs have exhibited remarkable proficiency in both switching performance and channel conductance, due to the efficient isolation of the writing and reading processes associated with channel conductance. This distinctive characteristic renders IGTs a preferred choice for serving as synaptic devices in the field of neural computing [46].

FGT can store electrons for a long time (around 10 years). Since then, the FG transistor as a memory device can used for the semiconductor industry commercially. This non-volatile memory device has many advantages such as long durability, larger storage capacity, and higher flexibility. The programmable synapses based on floating-gate technology prove highly advantageous for analog VLSI neural circuits. This is attributed to their robust long-term information storage capabilities, adaptability for neural reprogramming, and seamless integration with conventional VLSI circuits during the manufacturing process [146].

FeFET stands out with its diverse advantages, encompassing swift read/write speeds, elevated density, minimal power usage, non-destructive readout capability, random access convenience, and unparalleled endurance. These attributes set it apart from traditional mainstream flash memories [147]. The captivating feature of ferroelectric materials in neuromorphic applications arises from their bistable memory effect. Bistability manifests when a device exhibits two distinct resistive states, and notably, the device retains its state even after the applied bias is withdrawn.

XI. OPEN PROBLEMS, CHALLENGES, LIMITATIONS, CURRENT RESEARCH AREAS AND FUTURE PROSPECTS

Neuromorphic computing hardware using transistor-based resistive memory is challenged by the dynamic switching behavior of the nodes. A higher programming voltage for faster switching to programmable states demands a higher power consumption. The flow of current in a semiconductor material can be enhanced by temperature, which helps reduce the switching delay. An increase in temperature leads to increased power consumption and unreliable programming. From an architectural point of view, the physical, electromagnetic, and mutual inductance effects of different circuit components enhance the chances of losing signal integrity and signal attenuation. The effects of electrical and thermal noises are other factors to be considered.

Crossbar nodes consisting of transistors and resistive memory devices are used. CMOS transistors are used in applications of neuromorphic hardware based on Multiply and Accumulate (MAC)operations. The Multiply and Accumulate operation mimics the weighted summation in a neural network. In a neural network, weights are multiplied by the neural inputs, and these results are added. The summated result will fed to the activation function for further processing through the upcoming neural network layers. In a MAC operation using a memristive crossbar array, the neural network inputs are mapped to the input voltages, and the weight values are mapped to the conductance states of the memristor nodes. Memristor crossbar in which memristor nodes are arranged in a matrix form, the input voltages are fed to the rows, and the output currents measured from the columns represent the summation of products of input voltage and memristive conductance. Hence, the weighted summation of neural network and MAC operation in a memristive crossbar array is analogous. CMOS transistors are used in crossbar nodes, inputs, or outputs to select the particular device and to avoid the unintentional flow of current through an undesired path. The parasitic effects and leakage currents associated with CMOS transistors cause extra power consumption by affecting accurate crossbar computations. Parasitic capacitances are formed owing to the separation of mobile charge careers in different regions within the CMOS structure. These unwanted parasitic capacitances can be neglected when the CMOS circuit operates at low frequency. At low frequencies, the capacitative impedance is high (considered as infinity), acts as an open circuit, and does not affect the circuit. At high frequencies, they act as impedances and affect the transistor behavior. There is a chance of current leakage between the drain and the source of a MOSFET, even if it is off. This leakage current and unwanted parasitic effects cause high power loss. The current flowing through the unintentional circuit paths is called the sneak path current. The major aim of transistors is to reduce sneak path issues. However, practically, even though the transistors are in the cutoff region, there is a subthreshold current flow.

The nonidealities in the resistive memory devices in the crossbar node lead to relative current error at the output, relative crossbar leakage current, and performance deterioration [148], [149], [150], [151]. The crossbar leakage current increases with an increase in crossbar size even if nanometer-scale metal wires of negligible resistance are used. The equivalent capacitance of all resistive memory devices in a row of memristive crossbars is called the line capacitance. This line capacitance directly affects the dynamic behavior of the memristor. When the switching time of the memristors is low, the impact is not significant. However, this will lead to computational errors and power wastage. Improper packaging leads to deterioration of the performance of crossbars and improper utilization of power. A crossbar architecture consisting of transistors and resistive memory devices with fast switching exhibits the capacitative effect of bond wires, and self-inductive effects distort the transmitted signal. IR drop and wire resistance are the two other negative impacts of the sneak-path current [152], [153]. The current passing through the metal wires leads to a significant voltage drop in metal wires. When the crossbar size increased, this drop also increased because of increased metal resistance. Even without considering the sneak path current, an IR drop issue exists because the current passes through different resistances, producing a voltage drop. When the size of the crossbar increases, this IR drop issue also increases, leading to power wastage. Metal lines that are used to connect memristors offer resistance which is known as wire resistance. This wire resistance, resistors in the CMOS switches, and stray capacitors create RC delays along crossbar rows and columns. This reduces the read speed, particularly when the crossbar size increases. Parasitic effects lead to a large current error exhaustion of the power.

When the number of computational blocks increases, the power consumption also increases. The Input/Output blocks associated with memristive crossbars such as opamps, ADC, buffers, sense amplifiers, line selector switches, programming circuits, and multiplexers account for the majority of power usage. The opamp, or set of amplifiers, reads the column currents. To consume minimal power, some factors associated with the offset voltage/current, and delays should be minimal. The parasitic resistors and capacitors from the sense amplifiers cause a delay in the output reading. To address these issues, extra timing or memory control circuits are necessary, and these additional computational blocks enhance the power demand. To control the various switches in the crossbar architecture and peripheral circuits, different control circuits are required. Neural network applications in which multiple crossbars are used also require control circuits to access intermediate storage and multiplex signals between crossbars. The control circuit required additional power. The errors associated with the ADC/DAC in the peripheral circuits enhance the chance of current errors. Peripheral circuits are sensitive to voltage, process, temperature, and noise. Noise introduced in the peripheral circuits is passed from one crossbar array to another. The computational errors resulting from this were also passed. A modular crossbar approach/crossbar tiling is used to reduce the leakage current owing to the interconnection of resistance between rows and columns of crossbars and memristor variabilities. However, when the number of crossbar modules increases, the requirement for additional circuitry also increases to add up the output signals from different modules. Because of this, the power demand also increased due to this fact. Endurance and data retention are the two major aspects of memory devices used for storing data. Endurance denotes the capacity of the storage device to endure numerous write and erase cycles without compromising the integrity and reliability of the stored data. However, data retention signifies a storage device's capability to preserve stored data over time, ensuring that there is no degradation or loss of information integrity. High endurance and data retention values in synaptic applications ensure the longevity, stability, energy efficiency, accuracy, and reduced maintenance of neural networks, making them essential parameters for the successful implementation of neuromorphic computing systems. However, a higher value of endurance and data retention is very difficult to obtain due to variations in various factors present in the systems. Miniaturization often leads to reduced endurance and data retention because smaller devices tend to be more vulnerable to various forms of degradation, such as electromigration or defect. The large dynamic ranges, optimum noise, and high switching ratio may help achieve multi-gate states and optimize device variability. A lower operating voltage reduces the power consumption during the read/write operation. Unfortunately, temperature sensitivity and various technological constraints may degrade the overall performance. The use of organic materials in IGT devices makes the device unstable and limits its speed [154]. Organic materials have a lower charge mobility. The low ionic mobility of common electrolytes and poor reversibility of ion penetration affect device speed and endurance. These materials are extremely expensive and difficult to synthesize. Additional procedures are required to maintain a good quality. Moreover, organic materials are vulnerable to damage from moisture and insects, and are more susceptible to fading, further complicating their practical application in IGT devices. Devices are mostly affected by nonideal factors such as voltage variations and conductivity states. Careful selection and design of electrolytes and channel materials, along with engineering the ion electrolyte/channel interface, are key steps for optimizing the IGT performance.

TABLE 4. Commercialized crossbar MACs.

Chip	Contributors	Memory element	Architecture	Analog/Digital	Features
3D XPoint Memory	Intel & Micron	Phase Change Memory (PCM)	Cross-point architecture in which memory elements with selectors are stacked in three-dimension	Digital. Low Resistance State: logic 1 High Resistance State: logic 0	*Ovonic Threshold Switch as selector *PCM SET time: 25ns *PCM RESET current:10uA *PCM Endurance: 10^{12} cycles *Thousand times faster than existing memory technologies.
AI mixed signal chip	IBM's Albany NanoTech Complex	Phase Change Memory (PCM)	64 crossbar tiles, in which each tile has 256 rows and 256 columns	Analog	*14nm CMOS technology *Maximum throughput: 63.1 TOPS *Energy efficiency: 9.76 TOPS/W * High weight capacity, parallelism throughput and accuracy
MB85AS4MT	Fujitsu Semiconductor	Resistive Random Access Memory (ReRAM)	ReRAM cell array of size 524288 × 8	Digital	*Silicon gate CMOS process & resistance-variable memory process *Endurance : 1.2×10^6 *Retention of 10 years *Operating voltage :1.65V to 3.6V *Rewrite current: 1.3mA *Read-out current: 0.2mA

The construction of neural networks requires more synaptic connections. However, integrating these connections through three-terminal devices such as IGTs, FG transistors, and Fe FETs introduces complex wiring and routing challenges. Unlike their two-terminal counterparts, these three-terminal devices require intricate peripheral control circuits, which add complexity. These routing challenges render the design and fabrication of large-scale networks highly complex. Achieving large-scale integration of three-terminal devices for constructing artificial synaptic and neural networks is a long-term goal [155]. Integrating three-terminal devices seamlessly with existing semiconductor technology and complementary metal-oxide-semiconductor (CMOS) circuits is challenging. Ensuring compatibility and efficient communication between different components of a neuromorphic system is essential for its overall functionality. Manufacturing processes can introduce variability into the characteristics of three-terminal devices, thereby affecting their performance consistency. Achieving uniformity and reliability across a large number of devices poses significant challenges during the fabrication process. The resources and procedures required to design and fabricate an integrated circuit differ at different scales. Scaling down certain limits presents many associated challenges. From an architectural point of view, the geometrical mismatch between circuit components is a primary factor to be considered [156]. The number of neural network layers to be implemented in two dimensions is limited owing to area constraints, even when modular crossbars are used. The three-dimensional stacking of layers also presents geometrical challenges for fitting intermediate circuits.

XII. COMMERCIALIZED CROSSBAR MACS

Table 4 provides details regarding commercialized crossbar chips for Multiply and Accumulate (MAC)operations. Intel and Micron jointly developed a 3D XPoint memory storage technology to fill the gap between existing technologies, dynamic RAM and NAND flash. The technology contributors claim that 3D XPoint will be a thousand times faster, a thousand times more endurance, and ten times the storage capacity than existing memory technologies.

3D XPoint has a cross-point architecture based on the memory technology Phase Change Memory (PCM), which is transistorless. The selectors and memory cells are positioned at the intersection of the perpendicular wires. The cells are accessed by the current through the top and bottom wires. The stacking of 3D XPoint cells in three dimensions improves the storage density. A single data, either 0 or 1, is stored in each cell. Storage is based on modifying the resistance level of the cell. Each cell exhibits two resistance levels. The high resistance state represents zero, and the low resistance state represents 1. It is retained in the latest attained state owing to its non-volatile nature. The reading and writing operations are performed by varying the voltage supplied to the selectors. During writing operations, according to the voltage supplied, the selector is activated, which helps change the resistance levels of the memory devices. During the reading operation, another range of voltage is applied to check whether the cell is in a high resistance or low resistance state. Unlike the NAND flash, 3DXpoint can write data at the bit level [157].

The IBM research community introduced an energy-efficient analog AI mixed signal chip for different Deep Neural Network inference tasks [158]. IBM's Albany NanoTech Complex was used to fabricate the chip. It has 64 tiles/ analog in-memory computation core. Each core consists of a crossbar array of 256 rows and 256 columns. Each tile is integrated with a time-based analog-to-digital converter. Tiles are also associated with lightweight digital processing units that perform non-linear neuronal activation functions and scaling. Digital communication pathways occur at the chip interconnects of all tiles, and the global digital processing unit. By performing analog neural computing on CIFAR-10, a precision of 92.81% was obtained.

MB85AS4MT is a chip based on ReRAM contributed by Fujitsu semiconductors in a configuration of 524,288 words × 8 bits. To form the non-volatile memory cells, resistance-variable memory process and silicon gate CMOS process technologies are used. Memory cells used are capable of having 1.2 × 10^6 rewrites.

Crossbar chips of configurations 2 × 2 × 16, 4 × 4 × 8, 8 × 8 × 4, 16 × 16 × 2 and 32 × 32 × 1 had contributed by Knowm that can be used for research in neural network accelerators, in-memory computing and non-volatile memory controllers.

CrossBar's ReRAM technology can be integrated between two metal lines and crossbar arrays can integrate on CMOS logic wafers to build a 3D ReRAM storage chip. This helps to provide on-chip, non-volatile memory with more advantages than NAND solutions.

XIII. RECENT INDUSTRY AND RESEARCH PROSPECTS OF BEYOND CMOS DEVICES

According to the International Roadmap for Devices and Systems (IRDS) 2022 report, STT MRAM has been manufactured commercially of embedded and standalone flash-like applications. Prominent companies like TSMC, GlobalFoundries, and Samsung have announced the production of embedded MRAM due to its benefits, including Non-volatile, exceptional durability, scalable, energy-efficient, and requiring fewer masks than embedded flash. SIT MRAM consumes less power and minimizes the leakage current compared to flash memory. Standalone MRAM products are also available IoT and data center applications.

Switching in OxRAM using the 1T1R configuration was pioneered by Toshiba, Panasonic, and IMEC. Toshiba unveiled a 32 GB RRAM chip integrated with 24 nm CMOS technology. Panasonic and IMEC presented an encapsulated cell structure with an Ir/Ta_2O_5/$\mathrm{Ta}O_x$/TaN stack on a 2-Mbit chip at the 40 nm node. [159].

The floating gate, also known as a synaptic transistor, has applications in analog and mixed-signal contexts owing to its low input and output impedances. These characteristics contribute to minimizing the overhead of peripheral circuitry. By implementing NAND flash memory, a floating gate neuromorphic circuit can be designed for high density. It has been reported that mixed-signal neuromorphic circuits with industrial-grade SONOS floating gate devices exhibit better performance, particularly in terms of energy efficiency [160]. The semiconductor industry has already developed 3D NAND memory with a floating gate of 96 layers. projections suggest an increase of 512 layers to attain a density of 10 Tb/in^2 density. This advancement is crucial for accommodating the storage requirements of large-scale neuromorphic models [161].

IBM, Infineon, Samsung, Macronix, and other entities tested the PCM prototypes. Additionally, collaborations between Intel and STMicroelectronics, as well as with Samsung, have recently announced the production of PCM, further highlighting its potential to compete with other conventional memory devices because of its cost-effectiveness, high speed, high density, and substantial non-volatile storage capacity. The FeFET technology has been successfully implemented using conventional HKMG technology, allowing the manufacture of a 28 nm FeFET device. This achievement is attributed to the reduced number of masks required in the fabrication process compared with the embedded FLASH. FE-HfO_2 based FeFET delivers superior performance with a faster switching speed (100 ns), operating voltages ranging from 4 to 6 V, and impressive ten-year data continuation and high durability, reaching up to 10^{12} switching cycles [162].

XIV. CONCLUSION AND FUTURE OUTLOOK

This review provides a comprehensive analysis of the rapidly evolving field of neuromorphic computing, with a specific emphasis on the role of transistors in enabling efficient and brain-inspired computing. We explored various transistor technologies, their integration into neuromorphic architectures, and the influence of emerging materials on their performance. IGTs offer numerous benefits, including elevated transconductance, rapid speed, and the capability for individual gating. These features make them highly suitable for amplifying small signals and processing neural signals effectively. FGT find widespread commercial application in various memory storage technologies, due to their enduring long-term durability. Ferroelectric-gate transistors possess the benefit of operating at high speeds. They can be manufactured using CMOS technology; nevertheless, they encounter numerous challenges unlike traditional transistors. Materials for transistor technology in neural computing applications must exhibit low power consumption, high speed, and tunability. Ferroelectric materials provide a promising path for transistor technology enabling the development of neuromorphic systems that mimic the brain's dynamic and adaptable behavior. Additionally, emerging 2D materials like graphene and transition metal dichalcogenides offer exceptional conductivity and flexibility, making them suitable for designing energy-efficient and adaptable synaptic transistors.

The field of neuromorphic computing is about to witness significant advancements and innovations in upcoming years. Future expectations include the development of advanced transistor technologies, creation of hybrid neuromorphic systems, design of specialized hardware accelerators, refinement of neuromorphic algorithms, and the consideration of ethical and security implications. For realising fully connected neural networks, memristive crossbar topology is widely used. The sparse connectivity in many neural networks in real applications makes mapping to a crossbar structure difficult. Scaling is also a challenge in the case of larger neural networks. Hence, the memristive crossbars are combined with discrete synapse modules in hybrid neuromorphic computing systems. In addition, the practical implementation of neuromorphic hardware accelerators in real-world applications will become prominent, contributing to the development of smart and autonomous systems. Neuromorphic algorithms will evolve to fully perform the capabilities of emerging transistors, enabling efficient learning and decision-making processes.

REFERENCES

[1] C. D. Schuman, S. R. Kulkarni, M. Parsa, J. P. Mitchell, P. Date, and B. Kay, "Opportunities for neuromorphic computing algorithms and applications," *Nature Comput. Sci.*, vol. 2, no. 1, pp. 10–19, Jan. 2022.

[2] D. Strukov, G. Indiveri, J. Grollier, and S. Fusi, "Building brain-inspired computing," *Nature Commun.*, vol. 10, p. 4838, Oct. 2019.

[3] A. Sebastian, M. Le Gallo, R. Khaddam-Aljameh, and E. Eleftheriou, "Memory devices and applications for in-memory computing," *Nature Nanotechnol.*, vol. 15, no. 7, pp. 529–544, Jul. 2020.

[4] L. Chua, "Memristor—The missing circuit element," *IEEE Trans. Circuit Theory*, vol. CT-18, no. 5, pp. 507–519, Sep. 1971.

[5] L. Chua, G. C. Sirakoulis, and A. Adamatzky, *Handbook of Memristor Networks*. Cham, Switzerland: Springer, 2019.

[6] D. B. Strukov, G. S. Snider, D. R. Stewart, and R. S. Williams, "The missing memristor found," *Nature*, vol. 453, no. 7191, pp. 80–83, May 2008.

[7] T. R. Rajalekshmi, R. R. Das, C. Reghuvaran, and A. James, "Graphene-based RRAM devices for neural computing," *Frontiers Neurosci.*, vol. 17, Oct. 2023, Art. no. 1253075.

[8] R. Yuste, "From the neuron doctrine to neural networks," *Nature Rev. Neurosci.*, vol. 16, no. 8, pp. 487–497, Aug. 2015.

[9] J. Q. Yang, R. Wang, Y. Ren, J. Y. Mao, Z. P. Wang, Y. Zhou, and S. T. Han, "Neuromorphic engineering: From biological to spike-based hardware nervous systems," *Adv. Mater.*, vol. 32, no. 52, Dec. 2020, Art. no. 2003610.

[10] D. Graupe, *Principles of Artificial Neural Networks*, 7th ed. Singapore: World Scientific, Sep. 2013.

[11] A. Mehonic, A. Sebastian, B. Rajendran, O. Simeone, E. Vasilaki, and A. J. Kenyon, "Memristors—From in-memory computing, deep learning acceleration, and spiking neural networks to the future of neuromorphic and bio-inspired computing," *Adv. Intell. Syst.*, vol. 2, no. 11, Nov. 2020, Art. no. 2000085.

[12] R. B. Salazar, H. Ilatikhameneh, R. Rahman, G. Klimeck, and J. Appenzeller, "A predictive analytic model for high-performance tunneling field-effect transistors approaching non-equilibrium Green's function simulations," *J. Appl. Phys.*, vol. 118, no. 16, Oct. 2015, Art. no. 164305.

[13] Y.-B. Kim, "Challenges for nanoscale MOSFETs and emerging nanoelectronics," *Trans. Electr. Electron. Mater.*, vol. 11, no. 3, pp. 93–105, Jun. 2010.

[14] Y. Zhu, Y. Zhu, H. Mao, Y. He, S. Jiang, L. Zhu, C. Chen, C. Wan, and Q. Wan, "Recent advances in emerging neuromorphic computing and perception devices," *J. Phys. D, Appl. Phys.*, vol. 55, no. 5, Feb. 2022, Art. no. 053002.

[15] M. Davies, A. Wild, G. Orchard, Y. Sandamirskaya, G. A. F. Guerra, P. Joshi, P. Plank, and S. R. Risbud, "Advancing neuromorphic computing with loihi: A survey of results and outlook," *Proc. IEEE*, vol. 109, no. 5, pp. 911–934, May 2021.

[16] D. Marković, A. Mizrahi, D. Querlioz, and J. Grollier, "Physics for neuromorphic computing," *Nature Rev. Phys.*, vol. 2, no. 9, pp. 499–510, Jul. 2020.

[17] D. Chatterjee and A. Kottantharayil, "A CMOS compatible bulk FinFET-based ultra low energy leaky integrate and fire neuron for spiking neural networks," *IEEE Electron Device Lett.*, vol. 40, no. 8, pp. 1301–1304, Aug. 2019.

[18] A. Balaji, A. Das, Y. Wu, K. Huynh, F. G. Dell'Anna, G. Indiveri, J. L. Krichmar, N. D. Dutt, S. Schaafsma, and F. Catthoor, "Mapping spiking neural networks to neuromorphic hardware," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 28, no. 1, pp. 76–86, Jan. 2020.

[19] J. Zhao, "Conversion of whetstone trained spiking deep neural networks to spiking neural networks," Ph.D. thesis, Dept. Comput. Sci., Univ. Tennessee, Knoxville, TN, USA, 2019.

[20] Y. Li, Z. Wang, R. Midya, Q. Xia, and J. J. Yang, "Review of memristor devices in neuromorphic computing: Materials sciences and device challenges," *J. Phys. D, Appl. Phys.*, vol. 51, no. 50, Dec. 2018, Art. no. 503002.

[21] P. Sheridan and W. Lu, "Memristors and memristive devices for neuromorphic computing," in *Memristor Networks*. New York, NY, USA: Springer, 2014, pp. 129–149.

[22] T. Prodromakis and C. Toumazou, "A review on memristive devices and applications," in *Proc. 17th IEEE Int. Conf. Electron., Circuits Syst.*, Dec. 2010, pp. 934–937.

[23] I. Boybat, M. Le Gallo, S. R. Nandakumar, T. Moraitis, T. Parnell, T. Tuma, B. Rajendran, Y. Leblebici, A. Sebastian, and E. Eleftheriou, "Neuromorphic computing with multi-memristive synapses," *Nature Commun.*, vol. 9, no. 1, pp. 1–12, Jun. 2018.

[24] J. Chen, J. Li, Y. Li, and X. Miao, "Multiply accumulate operations in memristor crossbar arrays for analog computing," *J. Semiconductors*, vol. 42, no. 1, Jan. 2021, Art. no. 013104.

[25] C. Yakopcic, M. Z. Alom, and T. M. Taha, "Extremely parallel memristor crossbar architecture for convolutional neural network implementation," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, May 2017, pp. 1696–1703.

[26] J. A. Starzyk and Basawaraj, "Memristor crossbar architecture for synchronous neural networks," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 61, no. 8, pp. 2390–2401, Aug. 2014.

[27] Y. Li and K.-W. Ang, "Hardware implementation of neuromorphic computing using large-scale memristor crossbar arrays," *Adv. Intell. Syst.*, vol. 3, no. 1, Jan. 2021, Art. no. 2000137.

[28] O. Krestinskaya, B. Choubey, and A. P. James, "Memristive GAN in analog," *Sci. Rep.*, vol. 10, no. 1, p. 5838, Apr. 2020.

[29] J. S. Meena, S. M. Sze, U. Chand, and T.-Y. Tseng, "Overview of emerging nonvolatile memory technologies," *Nanosc. Res. Lett.*, vol. 9, no. 1, pp. 1–33, Dec. 2014.

[30] W. Banerjee, "Challenges and applications of emerging nonvolatile memory devices," *Electronics*, vol. 9, no. 6, p. 1029, Jun. 2020.

[31] T. Andre, S. M. Alam, D. Gogl, C. K. Subramanian, H. Lin, W. Meadows, X. Zhang, N. D. Rizzo, J. Janesky, D. Houssameddine, and J. M. Slaughter, "ST-MRAM fundamentals, challenges, and applications," in *Proc. IEEE Custom Integr. Circuits Conf.*, Sep. 2013, pp. 1–8.

[32] D. Apalkov, A. Khvalkovskiy, S. Watts, V. Nikitin, X. Tang, D. Lottis, K. Moon, X. Luo, E. Chen, A. Ong, A. Driskill-Smith, and M. Krounbi, "Spin-transfer torque magnetic random access memory (STT-MRAM)," *ACM J. Emerg. Technol. Comput. Syst. (JETC)*, vol. 9, no. 2, pp. 1–35, 2013.

[33] P. Gölitz and R. Hecker, "Turning potential into realities," *Chemphyschem, Eur. J. Chem. Phys. Phys. Chem.*, vol. 2, nos. 8–9, pp. 473–474, 2001.

[34] M. Riordan, L. Hoddeson, and C. Herring, "The invention of the transistor," *Rev. Mod. Phys.*, vol. 71, no. 2, p. S336, 1999.

[35] W. Shockley, "The path to the conception of the junction transistor," *IEEE Trans. Electron Devices*, vol. ED-23, no. 7, pp. 597–620, Jul. 1976.

[36] M. Grundmann, "Kramers–Kronig relations," in *The Physics of Semiconductors: An Introduction Including Nanophysics and Applications*. Berlin, Germany: Springer, 2010, pp. 775–776.

[37] M. M. Atalla, E. Tannenbaum, and E. J. Scheibner, "Stabilization of silicon surfaces by thermally grown oxides," *Bell Syst. Tech. J.*, vol. 38, no. 3, pp. 749–783, May 1959.

[38] R. G. Arns, "The other transistor: Early history of the metal-oxide semiconductor field-effect transistor," *Eng. Sci. Educ. J.*, vol. 7, no. 5, pp. 233–240, Oct. 1998.

[39] R. R. Das, S. Maity, A. Choudhury, A. Chakraborty, C. T. Bhunia, and P. P. Sahu, "Temperature-dependent short-channel parameters of FinFETs," *J. Comput. Electron.*, vol. 17, no. 3, pp. 1001–1012, Sep. 2018.

[40] R. R. Das, A. Chowdhury, A. Chakraborty, and S. Maity, "Impact of stress effect on triple material gate step-FinFET with DC and AC analysis," *Microsyst. Technol.*, vol. 26, no. 6, pp. 1813–1821, Jun. 2020.

[41] J. Bhardwaj, K. K. Gupta, and R. Gupta, "A review of emerging trends on water quality measurement sensors," in *Proc. Int. Conf. Technol. Sustain. Develop. (ICTSD)*, Feb. 2015, pp. 1–6.

[42] W. R. Curtice, "A MESFET model for use in the design of GaAs integrated circuits," *IEEE Trans. Microw. Theory Techn.*, vol. MTT-28, no. 5, pp. 448–456, May 1980.

[43] Y.-M. Lin, H.-Y. Chiu, K. A. Jenkins, D. B. Farmer, P. Avouris, and A. Valdes-Garcia, "Dual-gate graphene FETs with f_T of 50 GHz," *IEEE Electron Device Lett.*, vol. 31, no. 1, pp. 68–70, Jan. 2010.

[44] H. S. White, G. P. Kittlesen, and M. S. Wrighton, "Chemical derivatization of an array of three gold microelectrodes with polypyrrole: Fabrication of a molecule-based transistor," *J. Amer. Chem. Soc.*, vol. 106, no. 18, pp. 5375–5377, Sep. 1984.

[45] A. M. Deml, A. L. Bunge, M. A. Reznikov, A. Kolessov, and R. P. O'Hayre, "Progress toward a solid-state ionic field effect transistor," *J. Appl. Phys.*, vol. 111, no. 7, Apr. 2012, Art. no. 074511.

[46] X. Bu, H. Xu, D. Shang, Y. Li, H. Lv, and Q. Liu, "Ion-gated transistor: An enabler for sensing and computing integration," *Adv. Intell. Syst.*, vol. 2, no. 12, Dec. 2020, Art. no. 2000156.

[47] M. Takayanagi, T. Tsuchiya, W. Namiki, T. Higuchi, and K. Terabe, "Correlated metal $SrVO_3$ based all-solid-state redox transistors achieved by Li^+ or H^+ transport," *J. Phys. Soc. Jpn.*, vol. 87, no. 3, Mar. 2018, Art. no. 034802.

[48] T. Tsuchiya, K. Terabe, R. Yang, and M. Aono, "Nanoionic devices: Interface nanoarchitechtonics for physical property tuning and enhancement," *Jpn. J. Appl. Phys.*, vol. 55, no. 11, Nov. 2016, Art. no. 1102A4.

[49] J. Liu, Z. Qin, H. Gao, H. Dong, J. Zhu, and W. Hu, "Vertical organic field-effect transistors," *Adv. Funct. Mater.*, vol. 29, no. 17, 2019, Art. no. 1808453.

[50] W. Shi, Y. Guo, and Y. Liu, "When flexible organic Field-Effect transistors meet biomimetics: A prospective view of the Internet of Things," *Adv. Mater.*, vol. 32, no. 15, Apr. 2020, Art. no. 1901493.

[51] M. Acosta, N. Novak, V. Rojas, S. Patel, R. Vaish, J. Koruza, G. A. Rossetti, and J. Rödel, "$BaTiO_3$-based piezoelectrics: Fundamentals, current status, and perspectives," *Appl. Phys. Rev.*, vol. 4, no. 4, p. 41305, Dec. 2017.

[52] K. Kim and S. Lee, "Integration of lead zirconium titanate thin films for high density ferroelectric random access memory," *J. Appl. Phys.*, vol. 100, no. 5, Sep. 2006, Art. no. 051604.

[53] H.-T. Lue, C.-J. Wu, and T.-Y. Tseng, "Device modeling of ferroelectric memory field-effect transistor (FeMFET)," *IEEE Trans. Electron Devices*, vol. 49, no. 10, pp. 1790–1798, Oct. 2002.

[54] M. Jerry, S. Dutta, A. Kazemi, K. Ni, J. Zhang, P.-Y. Chen, P. Sharma, S. Yu, X. S. Hu, M. Niemier, and S. Datta, "A ferroelectric field effect transistor based synaptic weight cell," *J. Phys. D, Appl. Phys.*, vol. 51, no. 43, Oct. 2018, Art. no. 434001.

[55] M.-K. Kim and J.-S. Lee, "Ferroelectric analog synaptic transistors," *Nano Lett.*, vol. 19, no. 3, pp. 2044–2050, Mar. 2019.

[56] K. Toprasertpong, M. Takenaka, and S. Takagi, "Memory window in ferroelectric field-effect transistors: Analytical approach," *IEEE Trans. Electron Devices*, vol. 69, no. 12, pp. 7113–7119, Dec. 2022.

[57] D. Kahng and S. M. Sze, "A floating gate and its application to memory devices," *Bell Syst. Tech. J.*, vol. 46, no. 6, pp. 1288–1295, Jul. 1967.

[58] H. Chen, Y. Zhou, and S. Han, "Recent advances in metal nanoparticle-based floating gate memory," *Nano Select*, vol. 2, no. 7, pp. 1245–1265, Jul. 2021.

[59] E. M. Conwell, *High Field Transport in Semiconductor* (Solid State Physics). New York, NY, USA: Academic, 1967.

[60] L. Esaki, "Long journey into tunneling," *Science*, vol. 183, no. 4130, pp. 1149–1155, Mar. 1974.

[61] C. Hu and M. A. Lieberman, *Electronics Research Laboratory*, document Contract 49620(90-C):0029, Boiling Air Force Base, Washington, DC, USA, 1998.

[62] C. A.-P. de Araujo, J. D. Cuchiaro, L. D. Mcmillan, M. C. Scott, and J. F. Scott, "Fatigue-free ferroelectric capacitors with platinum electrodes," *Nature*, vol. 374, no. 6523, pp. 627–629, Apr. 1995.

[63] J. De Blauwe, "Nanocrystal nonvolatile memory devices," *IEEE Trans. Nanotechnol.*, vol. 1, no. 1, pp. 72–77, Mar. 2002.

[64] Z. Liu, F. Xue, Y. Su, Y. M. Lvov, and K. Varahramyan, "Memory effect of a polymer thin-film transistor with self-assembled gold nanoparticles in the gate dielectric," *IEEE Trans. Nanotechnol.*, vol. 5, no. 4, pp. 379–384, Jul. 2006.

[65] S.-W. Ryu, J.-W. Lee, J.-W. Han, S. Kim, and Y.-K. Choi, "Designed workfunction engineering of double-stacked metal nanocrystals for nonvolatile memory application," *IEEE Trans. Electron Devices*, vol. 56, no. 3, pp. 377–382, Mar. 2009.

[66] M.-F. Chang, P.-T. Lee, S. P. McAlister, and A. Chin, "A flexible organic pentacene nonvolatile memory based on high-k dielectric layers," *Appl. Phys. Lett.*, vol. 93, no. 23, p. 439, Dec. 2008.

[67] S. Kim, T. Todorov, M. Onen, T. Gokmen, D. Bishop, P. Solomon, K.-T. Lee, M. Copel, D. B. Farmer, J. A. Ott, T. Ando, H. Miyazoe, V. Narayanan, and J. Rozen, "Metal-oxide based, CMOS-compatible ECRAM for deep learning accelerator," in *IEDM Tech. Dig.*, Dec. 2019, pp. 35.7.1–35.7.4.

[68] J. Tang, D. Bishop, S. Kim, M. Copel, T. Gokmen, T. Todorov, S. Shin, K.-T. Lee, P. Solomon, K. Chan, W. Haensch, and J. Rozen, "ECRAM as scalable synaptic cell for high-speed, low-power neuromorphic computing," in *IEDM Tech. Dig.*, Dec. 2018, pp. 13.1.1–13.1.4.

[69] W. Wang et al., "Neuromorphic sensorimotor loop embodied by monolithically integrated, low-voltage, soft e-skin," *Science*, vol. 380, no. 6646, pp. 735–742, May 2023.

[70] A. Melianas, T. J. Quill, G. LeCroy, Y. Tuchman, H. V. Loo, S. T. Keene, A. Giovannitti, H. R. Lee, I. P. Maria, I. McCulloch, and A. Salleo, "Temperature-resilient solid-state organic artificial synapses for neuromorphic computing," *Sci. Adv.*, vol. 6, no. 27, Jul. 2020, Art. no. eabb2958.

[71] Y. van de Burgt, E. Lubberman, E. J. Fuller, S. T. Keene, G. C. Faria, S. Agarwal, M. J. Marinella, A. A. Talin, and A. Salleo, "A non-volatile organic electrochemical device as a low-voltage artificial synapse for neuromorphic computing," *Nature Mater.*, vol. 16, no. 4, pp. 414–418, Apr. 2017.

[72] S. Rashidi, M. Jalili, and H. Sarbazi-Azad, "Improving MLC PCM performance through relaxed write and read for intermediate resistance levels," *ACM Trans. Archit. Code Optim.*, vol. 15, no. 1, pp. 1–31, Mar. 2018.

[73] F. Rao, K. Ding, Y. Zhou, Y. Zheng, M. Xia, S. Lv, Z. Song, S. Feng, I. Ronneberger, R. Mazzarello, W. Zhang, and E. Ma, "Reducing the stochasticity of crystal nucleation to enable subnanosecond memory writing," *Science*, vol. 358, no. 6369, pp. 1423–1427, Dec. 2017.

[74] J. Liang, R. G. D. Jeyasingh, H.-Y. Chen, and H.-S. P. Wong, "A 1.4 μA reset current phase change memory cell with integrated carbon nanotube electrodes for cross-point memory application," in *Proc. Symp. VLSI Technol.-Dig. Tech. Papers*, Jun. 2011, pp. 100–101.

[75] I. S. Kim, S. L. Cho, D. H. Im, E. H. Cho, D. H. Kim, G. H. Oh, D. H. Ahn, S. O. Park, S. W. Nam, J. T. Moon, and C. H. Chung, "High performance PRAM cell scalable to sub-20 nm technology with below $4F^2$ cell size, extendable to DRAM applications," in *Proc. Symp. VLSI Technol.*, Jun. 2010, pp. 203–204.

[76] G. Navarro et al., "Trade-off between SET and data retention performance thanks to innovative materials for phase-change memory," in *IEDM Tech. Dig.*, Dec. 2013, pp. 21.5.1–21.5.4.

[77] B. Walters, M. V. Jacob, A. Amirsoleimani, and M. R. Azghadi, "A review of graphene-based memristive neuromorphic devices and circuits," *Adv. Intell. Syst.*, vol. 2, no. 10, 2023, Art. no. 2300136.

[78] B. J. Choi, A. C. Torrezan, J. P. Strachan, P. G. Kotula, A. J. Lohn, M. J. Marinella, Z. Li, R. S. Williams, and J. J. Yang, "High-speed and low-energy nitride memristors," *Adv. Funct. Mater.*, vol. 26, no. 29, pp. 5290–5296, Aug. 2016.

[79] I. Mihai Miron, G. Gaudin, S. Auffret, B. Rodmacq, A. Schuhl, S. Pizzini, J. Vogel, and P. Gambardella, "Current-driven spin torque induced by the Rashba effect in a ferromagnetic metal layer," *Nature Mater.*, vol. 9, no. 3, pp. 230–234, Mar. 2010.

[80] M.-J. Lee, C. B. Lee, D. Lee, S. R. Lee, M. Chang, J. H. Hur, Y.-B. Kim, C.-J. Kim, D. H. Seo, S. Seo, U.-I. Chung, I.-K. Yoo, and K. Kim, "A fast, high-endurance and scalable non-volatile memory device made from asymmetric Ta_2O_{5-x}/TaO_{2-x} bilayer structures," *Nature Mater.*, vol. 10, no. 8, pp. 625–630, Aug. 2011.

[81] H. Jiang, L. Han, P. Lin, Z. Wang, M. H. Jang, Q. Wu, M. Barnell, J. J. Yang, H. L. Xin, and Q. Xia, "Sub-10 nm Ta channel responsible for superior performance of a HfO_2 memristor," *Sci. Rep.*, vol. 6, no. 1, Jun. 2016, Art. no. 28525.

[82] A. Chanthbouala, A. Crassous, V. Garcia, K. Bouzehouane, S. Fusil, X. Moya, J. Allibe, B. Dlubak, J. Grollier, S. Xavier, C. Deranlot, A. Moshar, R. Proksch, N. D. Mathur, M. Bibes, and A. Barthélémy, "Solid-state memories based on ferroelectric tunnel junctions," *Nature Nanotechnol.*, vol. 7, no. 2, pp. 101–104, Feb. 2012.

[83] S. Boyn, S. Girod, V. Garcia, S. Fusil, S. Xavier, C. Deranlot, H. Yamada, C. Carrétéro, E. Jacquet, M. Bibes, A. Barthélémy, and J. Grollier, "High-performance ferroelectric memory based on fully patterned tunnel junctions," *Appl. Phys. Lett.*, vol. 104, no. 5, Feb. 2014, Art. no. 052909.

[84] H. Yamada, V. Garcia, S. Fusil, S. Boyn, M. Marinova, A. Gloter, S. Xavier, J. Grollier, E. Jacquets, C. Carrétéro, C. Deranlot, M. Bibes, and A. Barthélémy, "Giant electroresistance of super-tetragonal $BiFeO_3$-based ferroelectric tunnel junctions," *ACS Nano*, vol. 7, no. 6, pp. 5385–5390, Jun. 2013.

[85] J. Lenz, F. D. Giudice, F. R. Geisenhof, F. Winterer, and R. T. Weitz, "Vertical, electrolyte-gated organic transistors show continuous operation in the MA cm^{-2} regime and artificial synaptic behaviour," *Nat. Nanotechnol.*, vol. 14, pp. 579–585, Mar. 2019.

[86] C. S. Yang, D. S. Shang, N. Liu, G. Shi, X. Shen, R. C. Yu, Y. Q. Li, and Y. Sun, "A synaptic transistor based on quasi-2D molybdenum oxide," *Adv. Mater.*, vol. 29, no. 27, Jul. 2017, Art. no. 1700906.

[87] L. Q. Zhu, C. J. Wan, L. Q. Guo, Y. Shi, and Q. Wan, "Artificial synapse network on inorganic proton conductor for neuromorphic systems," *Nature Commun.*, vol. 5, no. 1, p. 3158, Jan. 2014.

[88] Y. van de Burgt, A. Melianas, S. T. Keene, G. Malliaras, and A. Salleo, "Organic electronics for neuromorphic computing," *Nature Electron.*, vol. 1, pp. 386–397, Jul. 2018.

[89] Y. Chen, H. Yu, J. Gong, M. Ma, H. Han, H. Wei, and W. Xu, "Artificial synapses based on nanomaterials," *Nanotechnology*, vol. 30, no. 1, Jan. 2019, Art. no. 012001.

[90] Y. Kim, A. Chortos, W. Xu, Y. Liu, J. Y. Oh, D. Son, J. Kang, A. M. Foudeh, C. Zhu, Y. Lee, S. Niu, J. Liu, R. Pfattner, Z. Bao, and T.-W. Lee, "A bioinspired flexible organic artificial afferent nerve," *Science*, vol. 360, no. 6392, pp. 998–1003, Jun. 2018.

[91] E. J. Fuller, S. T. Keene, A. Melianas, Z. Wang, S. Agarwal, Y. Li, Y. Tuchman, C. D. James, M. J. Marinella, A. Salleo, and A. A. Talin, "Parallel programming of an ionic floating-gate memory array for scalable neuromorphic computing," *Science*, vol. 364, no. 6440, pp. 570–574, 2019.

[92] S. Wang, L. Liu, L. Gan, H. Chen, X. Hou, Y. Ding, S. Ma, D. W. Zhang, and P. Zhou, "Two-dimensional ferroelectric channel transistors integrating ultra-fast memory and neural computing," *Nature Commun.*, vol. 12, no. 1, pp. 1–9, Jan. 2021.

[93] T. R. Rajalekshmi, R. R. Das, R. Chithra, and A. James, "Graphene-based RRAM devices for neural computing," 2023, *arXiv:2308.02767*.

[94] Y.-S. Liu and P. Su, "Comparison of 2-D MoS$_2$ and Si ferroelectric FET nonvolatile memories considering the trapped-charge-induced variability," *IEEE Trans. Electron Devices*, vol. 69, no. 5, pp. 2738–2740, May 2022.

[95] M. Hassanpour Amiri, J. Heidler, K. Müllen, and K. Asadi, "Design rules for memories based on graphene ferroelectric field-effect transistors," *ACS Appl. Electron. Mater.*, vol. 2, no. 1, pp. 2–8, Jan. 2020.

[96] X. Jiang, X. Hu, J. Bian, K. Zhang, L. Chen, H. Zhu, Q. Sun, and D. W. Zhang, "Ferroelectric field-effect transistors based on WSe$_2$/CuInP$_2$S$_6$ heterostructures for memory applications," *ACS Appl. Electron. Mater.*, vol. 3, no. 11, pp. 4711–4717, 2021.

[97] M. Liu, T. Liao, Z. Sun, Y. Gu, and L. Kou, "2D ferroelectric devices: Working principles and research progress," *Phys. Chem. Chem. Phys.*, vol. 23, no. 38, pp. 21376–21384, 2021.

[98] S. Wan, Y. Li, W. Li, X. Mao, C. Wang, C. Chen, J. Dong, A. Nie, J. Xiang, Z. Liu, W. Zhu, and H. Zeng, "Nonvolatile ferroelectric memory effect in ultrathin α-In$_2$Se$_3$," *Adv. Funct. Mater.*, vol. 29, no. 20, May 2019, Art. no. 1808606.

[99] M. Si, P.-Y. Liao, G. Qiu, Y. Duan, and P. D. Ye, "Ferroelectric field-effect transistors based on MoS$_2$ and CuInP$_2$S$_6$ two-dimensional van der Waals heterostructure," *ACS Nano*, vol. 12, no. 7, pp. 6700–6705, Jul. 2018.

[100] T. Paul, T. Ahmed, K. K. Tiwari, C. S. Thakur, and A. Ghosh, "A high-performance MoS$_2$ synaptic device with floating gate engineering for neuromorphic computing," *2D Mater.*, vol. 6, no. 4, Jul. 2019, Art. no. 045008.

[101] C. Riggert, M. Ziegler, D. Schroeder, W. H. Krautschneider, and H. Kohlstedt, "MemFlash device: Floating gate transistors as memristive devices for neuromorphic computing," *Semiconductor Sci. Technol.*, vol. 29, no. 10, Oct. 2014, Art. no. 104011.

[102] Y. He, R. Liu, S. Jiang, C. Chen, L. Zhu, Y. Shi, and Q. Wan, "IGZO-based floating-gate synaptic transistors for neuromorphic computing," *J. Phys. D, Appl. Phys.*, vol. 53, no. 21, May 2020, Art. no. 215106.

[103] M.-S. Kim, J.-K. Kim, G.-J. Yun, J.-M. Yu, J.-K. Han, J.-W. Lee, S. Seo, S. Choi, and Y.-K. Choi, "An overturned charge injection synaptic transistor with a floating-gate for neuromorphic hardware computing," *IEEE Electron Device Lett.*, vol. 43, no. 9, pp. 1440–1443, Sep. 2022.

[104] A. Lahgere, "Design of leaky integrate and fire neuron for spiking neural networks using trench bipolar I-MOS," *IEEE Trans. Nanotechnol.*, vol. 22, pp. 260–265, 2023.

[105] K. Kim, C. Chen, Q. Truong, A. M. Shen, and Y. Chen, "A carbon nanotube synapse with dynamic logic and learning," *Adv. Mater.*, vol. 25, no. 12, pp. 1693–1698, Mar. 2013.

[106] Y. He, S. Nie, R. Liu, S. Jiang, Y. Shi, and Q. Wan, "Spatiotemporal information processing emulated by multiterminal neuro-transistor networks," *Adv. Mater.*, vol. 31, no. 21, May 2019, Art. no. 1900903.

[107] S. Jang, S. Jang, E.-H. Lee, M. Kang, G. Wang, and T.-W. Kim, "Ultrathin conformable organic artificial synapse for wearable intelligent device applications," *ACS Appl. Mater. Interfaces*, vol. 11, no. 1, pp. 1071–1080, Jan. 2019.

[108] C.-P. Chou, Y.-X. Lin, Y.-K. Huang, C.-Y. Chan, and Y.-H. Wu, "Junctionless poly-GeSn ferroelectric thin-film transistors with improved reliability by interface engineering for neuromorphic computing," *ACS Appl. Mater. Interfaces*, vol. 12, no. 1, pp. 1014–1023, Jan. 2020.

[109] M. A. Rodder, S. Vasishta, and A. Dodabalapur, "Double-gate MoS$_2$ field-effect transistor with a multilayer graphene floating gate: A versatile device for logic, memory, and synaptic applications," *ACS Appl. Mater. Interfaces*, vol. 12, no. 30, pp. 33926–33933, Jul. 2020.

[110] S. Kim, B. Choi, M. Lim, J. Yoon, J. Lee, H.-D. Kim, and S.-J. Choi, "Pattern recognition using carbon nanotube synaptic transistors with an adjustable weight update protocol," *ACS Nano*, vol. 11, no. 3, pp. 2814–2822, Mar. 2017.

[111] W. Liu and M. Asheghi, "Thermal conductivity measurements of ultra-thin single crystal silicon layers," *J. Heat Transf.*, vol. 128, no. 1, pp. 75–83, Jan. 2006.

[112] J. J. Petkowski, W. Bains, and S. Seager, "On the potential of silicon as a building block for life," *Life*, vol. 10, no. 6, p. 84, Jun. 2020.

[113] Y. Xiu, Y. Liu, D. W. Hess, and C. P. Wong, "Mechanically robust superhydrophobicity on hierarchically structured Si surfaces," *Nanotechnology*, vol. 21, no. 15, Apr. 2010, Art. no. 155705.

[114] S. Wagner, H. Gleskova, I. C. Cheng, and M. Wu, "Silicon for thin-film transistors," *Thin Solid Films*, vol. 430, nos. 1–2, pp. 15–19, Apr. 2003.

[115] L. Guo, P. R. Krauss, and S. Y. Chou, "Nanoscale silicon field effect transistors fabricated using imprint lithography," *Appl. Phys. Lett.*, vol. 71, no. 13, pp. 1881–1883, Sep. 1997.

[116] Z. Liu, L. Deng, and B. Peng, "Ferromagnetic and ferroelectric two-dimensional materials for memory application," *Nano Res.*, vol. 14, no. 6, pp. 1802–1813, Jun. 2021.

[117] J. Hoffman, X. Pan, J. W. Reiner, F. J. Walker, J. P. Han, C. H. Ahn, and T. P. Ma, "Ferroelectric field effect transistors for memory applications," *Adv. Mater.*, vol. 22, nos. 26–27, pp. 2957–2961, Jul. 2010.

[118] N. Nuraje and K. Su, "Perovskite ferroelectric nanomaterials," *Nanoscale*, vol. 5, no. 19, pp. 8752–8780, 2013.

[119] T. R. Rajalekshmi, V. Mishra, T. Dixit, P. R. Sagdeo, M. S. R. Rao, and K. Sethupathi, "Study of energy gaps and their temperature-dependent modulation in LaCrO$_3$: A theoretical and experimental approach," *J. Appl. Phys.*, vol. 133, no. 23, Jun. 2023, Art. no. 233104.

[120] T. R. Rajalekshmi, T. Dixit, M. S. R. Rao, and K. Sethupathi, "Pair-emission-induced near-infrared lasing from ceramic Ga:LaCrO$_3$ microcrystals at room temperature," *Phys. Status Solidi (RRL)*, vol. 15, no. 4, Apr. 2021, Art. no. 2000519.

[121] T. R. Rajalekshmi, V. Mishra, T. Dixit, M. Miryala, M. S. R. Rao, and K. Sethupathi, "Near white light and near-infrared luminescence in perovskite Ga:LaCrO$_3$," *Scripta Mater.*, vol. 210, Mar. 2022, Art. no. 114449.

[122] P. S. Halasyamani and K. R. Poeppelmeier, "Noncentrosymmetric oxides," *Chem. Mater.*, vol. 10, no. 10, pp. 2753–2769, Oct. 1998.

[123] M. Stewart, M. Cain, and D. Hall, "Ferroelectric hysteresis measurement and analysis," Nat. Phys. Lab., Teddington, MX, USA, Tech. Rep. NPL Report CMMT(A) 152, 1999.

[124] S. Han, Y. Zhou, and V. A. L. Roy, "Towards the development of flexible non-volatile memories," *Adv. Mater.*, vol. 25, no. 38, pp. 5425–5449, Oct. 2013.

[125] O. Auciello, J. F. Scott, and R. Ramesh, "The physics of ferroelectric memories," *Phys. Today*, vol. 51, no. 7, pp. 22–27, Jul. 1998.

[126] A. J. Lovinger, "Ferroelectric polymers," *Science*, vol. 220, no. 4602, pp. 1115–1121, Jun. 1983.

[127] R. Ramesh, T. Sands, V. G. Keramidas, and D. K. Fork, "Epitaxial ferroelectric thin films for memory applications," *Mater. Sci. Eng. B*, vol. 22, nos. 2–3, pp. 283–289, Jan. 1994.

[128] S. Oh, H. Hwang, and I. K. Yoo, "Ferroelectric materials for neuromorphic computing," *APL Mater.*, vol. 7, no. 9, Sep. 2019, Art. no. 091109.

[129] F. Xue, X. He, Y. Ma, D. Zheng, C. Zhang, L.-J. Li, J.-H. He, B. Yu, and X. Zhang, "Unraveling the origin of ferroelectric resistance switching through the interfacial engineering of layered ferroelectric-metal junctions," *Nature Commun.*, vol. 12, no. 1, pp. 1–8, Dec. 2021.

[130] E. Bersch, S. Rangan, R. A. Bartynski, E. Garfunkel, and E. Vescovo, "Band offsets of ultrathin high-*k* oxide films with Si," *Phys. Rev. B, Condens. Matter*, vol. 78, no. 8, Aug. 2008, Art. no. 085114.

[131] J. Robertson, "High dielectric constant gate oxides for metal oxide Si transistors," *Rep. Prog. Phys.*, vol. 69, no. 2, pp. 327–396, Feb. 2006.

[132] J. Robertson, "High dielectric constant oxides," *Eur. Phys. J. Appl. Phys.*, vol. 28, no. 3, pp. 265–291, Dec. 2004.

[133] F. Wu, H. Tian, Y. Shen, Z. Hou, J. Ren, G. Gou, Y. Sun, Y. Yang, and T.-L. Ren, "Vertical MoS_2 transistors with sub-1-nm gate lengths," *Nature*, vol. 603, no. 7900, pp. 259–264, Mar. 2022.

[134] Y.-S. Lin, R. Puthenkovilakam, and J. P. Chang, "Dielectric property and thermal stability of HfO_2 on silicon," *Appl. Phys. Lett.*, vol. 81, no. 11, pp. 2041–2043, Sep. 2002.

[135] A. Abderrahmane, C. Woo, and P. J. Ko, "Tunable optoelectronic properties of a two-dimensional graphene/α-In_2Se_3/graphene-based ferroelectric semiconductor field-effect transistor," *Res. Square*, vol. 32, pp. 20252–20258, Mar. 2021.

[136] D. Wu, A. J. Pak, Y. Liu, Y. Zhou, X. Wu, Y. Zhu, M. Lin, Y. Han, Y. Ren, H. Peng, Y.-H. Tsai, G. S. Hwang, and K. Lai, "Thickness-dependent dielectric constant of few-layer In_2Se_3 nanoflakes," *Nano Lett.*, vol. 15, no. 12, pp. 8136–8140, Dec. 2015.

[137] B. Arnaud, S. Lebègue, P. Rabiller, and M. Alouani, "Huge excitonic effects in layered hexagonal boron nitride," *Phys. Rev. Lett.*, vol. 96, no. 2, Jan. 2006, Art. no. 026402.

[138] S. Haastrup, M. Strange, M. Pandey, T. Deilmann, P. S. Schmidt, N. F. Hinsche, M. N. Gjerding, D. Torelli, P. M. Larsen, A. C. Riis-Jensen, J. Gath, K. W. Jacobsen, J. J. Mortensen, T. Olsen, and K. S. Thygesen, "The computational 2D materials database: High-throughput modeling and discovery of atomically thin crystals," *2D Mater.*, vol. 5, no. 4, Sep. 2018, Art. no. 042002.

[139] R. Geick, C. H. Perry, and G. Rupprecht, "Normal modes in hexagonal boron nitride," *Phys. Rev.*, vol. 146, no. 2, pp. 543–547, Jun. 1966.

[140] S. L. Howell, D. Jariwala, C.-C. Wu, K.-S. Chen, V. K. Sangwan, J. Kang, T. J. Marks, M. C. Hersam, and L. J. Lauhon, "Investigation of band-offsets at monolayer–multilayer MoS_2 junctions by scanning photocurrent microscopy," *Nano Lett.*, vol. 15, no. 4, pp. 2278–2284, Apr. 2015.

[141] M. Lederer, T. Kämpfe, T. Ali, F. Müller, R. Olivo, R. Hoffmann, N. Laleni, and K. Seidel, "Ferroelectric field effect transistors as a synapse for neuromorphic application," *IEEE Trans. Electron Devices*, vol. 68, no. 5, pp. 2295–2300, May 2021.

[142] T. Knobloch, "On the electrical stability of 2D material-based field-effect transistors," Ph.D. dissertation, Faculty Elect. Eng. Inf. Technol., TU Wien, Vienna, Austria, 2022.

[143] E. G. Marin, M. Perucchini, D. Marian, G. Iannaccone, and G. Fiori, "Modeling of electron devices based on 2-D materials," *IEEE Trans. Electron Devices*, vol. 65, no. 10, pp. 4167–4179, Oct. 2018.

[144] G. D. Spyropoulos, J. N. Gelinas, and D. Khodagholy, "Internal ion-gated organic electrochemical transistor: A building block for integrated bioelectronics," *Sci. Adv.*, vol. 5, no. 2, Feb. 2019, Art. no. eaau7378.

[145] K. Baeg and J. Lee, "Flexible electronic systems on plastic substrates and textiles for smart wearable technologies," *Adv. Mater. Technol.*, vol. 5, no. 7, Jul. 2020, Art. no. 2000071.

[146] B. W. Lee, B. J. Sheu, and H. Yang, "Analog floating-gate synapses for general-purpose VLSI neural computation," *IEEE Trans. Circuits Syst.*, vol. 38, no. 6, pp. 654–658, Jun. 1991.

[147] J. Ajayan, P. Mohankumar, D. Nirmal, L. M. I. L. Joseph, S. Bhattacharya, S. Sreejith, S. Kollem, S. Rebelli, S. Tayal, and B. Mounika, "Ferroelectric field effect transistors (FeFETs): Advancements, challenges and exciting prospects for next generation non-volatile memory (NVM) applications," *Mater. Today Commun.*, vol. 35, Jun. 2023, Art. no. 105591.

[148] Y. Jeong, M. A. Zidan, and W. D. Lu, "Parasitic effect analysis in memristor-array-based neuromorphic systems," *IEEE Trans. Nanotechnol.*, vol. 17, no. 1, pp. 184–193, Jan. 2018.

[149] O. Krestinskaya, A. Irmanova, and A. P. James, "Memristive non-idealities: Is there any practical implications for designing neural network chips?" in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, May 2019, pp. 1–5.

[150] F. Gül, "Addressing the sneak-path problem in crossbar RRAM devices using memristor-based one Schottky diode-one resistor array," *Results Phys.*, vol. 12, pp. 1091–1096, Mar. 2019.

[151] R. Naous, M. A. Zidan, A. Sultan-Salem, and K. N. Salama, "Memristor based crossbar memory array sneak path estimation," in *Proc. 14th Int. Workshop Cellular Nanosc. Netw. Appl. (CNNA)*, Jul. 2014, pp. 1–2.

[152] Y. Cassuto, S. Kvatinsky, and E. Yaakobi, "Sneak-path constraints in memristor crossbar arrays," in *Proc. IEEE Int. Symp. Inf. Theory*, Jul. 2013, pp. 156–160.

[153] L. Shi, G. Zheng, B. Tian, B. Dkhil, and C. Duan, "Research progress on solutions to the sneak path issue in memristor crossbar arrays," *Nanosc. Adv.*, vol. 2, no. 5, pp. 1811–1827, 2020.

[154] C. Cea, Z. Zhao, D. J. Wisniewski, G. D. Spyropoulos, A. Polyravas, J. N. Gelinas, and D. Khodagholy, "Integrated internal ion-gated organic electrochemical transistors for stand-alone conformable bioelectronics," *Nature Mater.*, vol. 22, no. 10, pp. 1227–1235, Oct. 2023.

[155] C. Sun, X. Liu, Q. Jiang, X. Ye, X. Zhu, and R.-W. Li, "Emerging electrolyte-gated transistors for neuromorphic perception," *Sci. Technol. Adv. Mater.*, vol. 24, no. 1, Dec. 2023, Art. no. 2162325.

[156] I. Vourkas, D. Stathis, G. Ch. Sirakoulis, and S. Hamdioui, "Alternative architectures toward reliable memristive crossbar memories," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 24, no. 1, pp. 206–217, Jan. 2016.

[157] M. Zabihi, S. Resch, H. Cilasun, Z. I. Chowdhury, Z. Zhao, U. R. Karpuzcu, J.-P. Wang, and S. S. Sapatnekar, "Exploring the feasibility of using 3-D XPoint as an in-memory computing accelerator," *IEEE J. Explor. Solid-State Comput. Devices Circuits*, vol. 7, pp. 88–96, 2021.

[158] M. Le Gallo et al., "A 64-core mixed-signal in-memory compute chip based on phase-change memory for deep neural network inference," *Nature Electron.*, vol. 6, no. 9, pp. 680–693, Aug. 2023.

[159] T.-Y. Liu et al., "A 130.7 mm^2 2-layer 32 Gb ReRAM memory device in 24 nm technology," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, San Francisco, CA, USA, Feb. 2013, pp. 140–153.

[160] L. Fick, D. Blaauw, D. Sylvester, S. Skrzyniarz, M. Parikh, and D. Fick, "Analog in-memory subthreshold deep neural network accelerator," in *Proc. IEEE Custom Integr. Circuits Conf. (CICC)*, Apr. 2017, pp. 1–4.

[161] C. Kim et al., "A 512-Gb 3-b/cell 64-stacked WL 3-D-NAND flash memory," *IEEE J. Solid-State Circuits*, vol. 53, no. 1, pp. 124–133, Jan. 2018.

[162] J. Müller, E. Yurchuk, T. Schlösser, J. Paul, R. Hoffmann, S. Müller, D. Martin, S. Slesazeck, P. Polakowski, J. Sundqvist, M. Czernohorsky, K. Seidel, P. Kücher, R. Boschke, M. Trentzsch, K. Gebauer, U. Schröder, and T. Mikolajick, "Ferroelectricity in HfO_2 enables nonvolatile data storage in 28 nm HKMG," in *Proc. Symp. VLSI Technol. (VLSIT)*, Jun. 2012, pp. 25–26.

RINKU RANI DAS received the B.E. degree in electronics and telecommunication engineering from Tripura Institute of Technology College, Tripura, India, in 2013, the M.Tech. degree in mobile communication and computing from the National Institute of Technology, Arunachal Pradesh, India, in 2017, and the Ph.D. degree in electronics and communication engineering from the National Institute of Technology, Agartala, Tripura, in 2023. She is currently an Electronics Design Engineer with the School of Electronics Systems and Automations, Digital University Kerala, India. Her research interests include semiconductor devices, such as FinFET, multi-bridge channel FET, and sensor.

T. R. RAJALEKSHMI received the bachelor's degree in physics from Kerala University, the master's degree in physics from the National Institute of Technology, Calicut, and the Ph.D. degree in physics from Indian Institute of Technology Madras. She is currently a Postdoctoral Fellow with Digital University Kerala. She has the work experience as a Project Fellow with the Space Physics Laboratory, Vikram Sarabhai Space Centre, during the master's degree. She was a recipient of Scholarship of Higher Education (SHE), awarded by Kerala State Higher Education Council (2010–2015), and also qualified National Eligibility Test for lectureship with Junior Research Fellowship in physical sciences. She is a fellow of Sakura Science Program, Japan, and also selected for the JASSO Fellowship for the Research Program with SIT, Japan. Her research interests include the study of transition metal oxides in bulk and thin film, 2D materials, graphene-based sensors, composite materials, and memory devices.

SRUTHI PALLATHUVALAPPIL (Graduate Student Member, IEEE) received the bachelor's degree in electronics and communication, in 2014, and the Master of Technology degree in embedded systems, in 2017. She is currently pursuing the Ph.D. degree with the School of Electronics Systems and Automation, Digital University Kerala. She is also involved in a few projects related to hardware-based low power memristive network implementation. Her research interests include memristive analog circuits, multi-bit logic memories, 3D integration, neuromorphic computing systems, and low-power resistive memory networks for AI.

ALEX JAMES (Senior Member, IEEE) received the Ph.D. degree from Griffith University, QLD, Australia. He is currently a Professor and the Dean (Academic) with Kerala University of Digital Sciences, Innovation and Technology (aka Digital University Kerala). He is also the Professor-in-Charge of the Maker Village, a Chief Investigator of the Centre for excellence in Intelligent IoT Sensors, and the Company Director of India Innovation Centre for Graphene. He is also the CTO of India Graphene Engineering and Innovation Centre. His research interests include AI—neuromorphic systems (software and hardware), VLSI, and image processing. He is the Creator of "Kairali AI processors," which is a low-power AI hardware and founded the startup Keralatoys. He is a member of IEEE CASS TC on Nonlinear Circuits and Systems, IEEE CTSoc TC on Quantum in Consumer Technology (QCT), TC on Machine learning, Deep learning and AI in CE (MDA), IEEE CASS TC on Cellular Nanoscale Networks and Memristor Array Computing (CNN-MAC), and IEEE CASS SIG on AgriElectronics. He is a Life Member of ACM, a Senior Fellow of HEA, a fellow of British Computer Society (FBCS), and a fellow of IET (FIET). He was awarded IEEE Outstanding researcher by IEEE Kerala Section for 2022, Kairali Scientist Award (Kairali Gaveshana Puraskaram) for Physical Science, in 2021 and 2022, and Best Associate Editor for TCAS1, in 2021. He was the Founding Chair of IEEE CASS Kerala Chapter, a member of IET Vision and Imaging Network, and currently a member of BCS' Fellows Technical Advisory Group (F-TAG). He was an Editorial Board Member of *Information Fusion* (2010–2014) and IEEE TRANSACTIONS ON CIRCUITS AND SYSTEM—I: REGULAR PAPERS (2018–2023); and currently serving as an Associate Editor for IEEE ACCESS, since 2017, *Frontiers in Neuroscience*, since 2022, IEEE OPEN JOURNAL OF CIRCUITS AND SYSTEMS, since 2022, and IEEE TRANSACTIONS ON BIOMEDICAL CIRCUITS AND SYSTEMS, since 2024. He has been the Associate Editor-in-Chief of IEEE OPEN JOURNAL OF CIRCUITS AND SYSTEMS, since 2024.