The Roadmap to Realize Memristive Three-Dimensional Neuromorphic Computing System

Hongyu An, Kangjun Bai and Yang Yi

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.78986

Abstract

Neuromorphic computing, an emerging non-von Neumann computing mimicking the physical structure and signal processing technique of mammalian brains, potentially achieves the same level of computing and power efficiencies of mammalian brains. This chapter will discuss the state-of-the-art research trend on neuromorphic computing with memristors as electronic synapses. Furthermore, a novel three-dimensional (3D) neuromorphic computing architecture combining memristor and monolithic 3D integration technology would be introduced; such computing architecture has capabilities to reduce the system power consumption, provide high connectivity, resolve the routing congestion issues, and offer the massively parallel data processing. Moreover, the design methodology of applying the capacitance formed by the through-silicon vias (TSVs) to generate a membrane potential in 3D neuromorphic computing system would be discussed in this chapter.

Keywords: memristor, synapse, three-dimensional integrated circuit, neuromorphic computing, analog/mixed-signal circuit design, monolithic 3D integration

1. Introduction

The continued success of the development in the modern von Neumann computing system was firstly enabled by the increment of the transistor integration density, followed by the multicore computing architecture. However, hindered by the fabrication process and size incompatibility between technologies of the complementary metal-oxide-semiconductor (CMOS) and the memory, central computing units (CPUs) and memory are located separately in resulting that the communication bus is inevitable. This communication bus becomes an energetic and speed bottleneck in this architecture. Furthermore, the transistor size shrinking

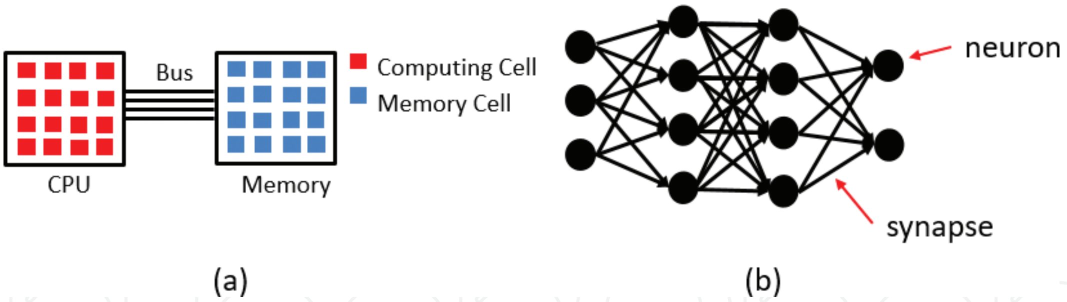

Figure 1. Architectures of (a) von Neumann computing system and (b) neuromorphic computing system.

trend is even harder to catch Moore's prediction due to the physical limitations [1]. As the density of data continuously escalates, extracting valuable information becomes computationally expensive, even for supercomputers. Meanwhile, the amount of energy required for supercomputers poses doubt on whether the increased performance is affordable.

On the other hand, as human beings, our brains have capabilities of learning and analyzing surrounding information with merely 20 W of power consumption [2]. Inspired by the working mechanism of the nervous system, the performance development of the computing system has led to a novel nontraditional computing architecture, namely, the neuromorphic computing system. The neuromorphic computing system was proposed by Carver Mead in the 1980s to mimic the mammalian neurology using the very-large-scaled-integrated (VLSI) circuit [3]. **Figure 1** illustrates the difference between the von Neumann architecture and the neuromorphic computing system. As powerful as the brain, the neuromorphic computing system potentially solves computing-intensive tasks that are only handled by the human brains before. These multifaceted tasks include speech recognition [4–6], character recognition [7, 8], grammar modeling [9], noise modeling [10], as well as the generation and prediction of chaotic time series [11, 12], etc. However, state-of-the-art neuromorphic chips with the traditional CMOS technology and the two-dimensional (2D) design methodology cannot meet the energetic and speed requirements at large-scale neuron and synapse realization [13–17]. In order to address this issue, recently, a three-dimensional (3D) neuromorphic computing architecture combining the memristors as electronic synapses is proposed and investigated [18–20].

This chapter is organized as follows, Section 2 introduces the background information of the neuromorphic computing, Section 3 discusses various neural models and their corresponding hardware implementations, Section 4 describes the biological reasons for employing memristive devices as electronic synapses, Section 5 illustrates the proposed 3D neuromorphic computing architecture, and at last, Section 6 draws some conclusions.

2. Neuromorphic computing

The digital computer based on the von Neumann architecture has powered our society for more than 40 years with its constant increment on computing capability. **Figure 2** shows the diagram of the typical von Neumann architecture. In this architecture, central computing

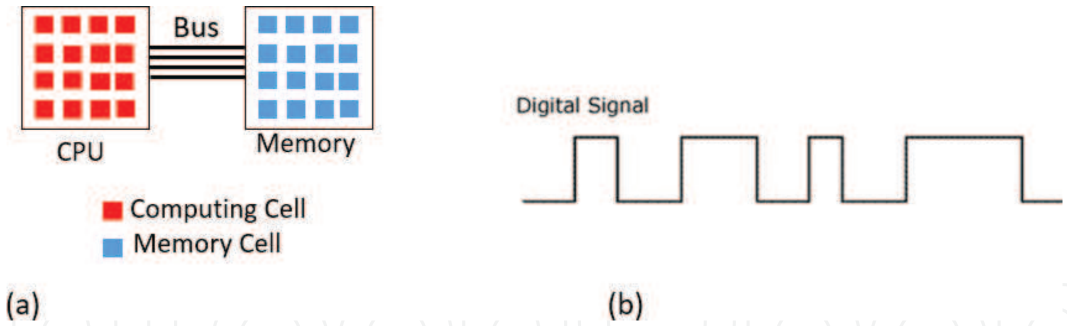

Figure 2. (a) The von Neumann architecture, (b) digital signal in computer.

units (CPUs) and memory units are physically separated at different locations due to their incompatibilities of fabrication process and size. A communication bus is used for the data transferring between them. In order to perform the Boolean algebra and arithmetic, the data stored in the memory need to be retrieved from the memory to CPU and be transferred back to memory after computing. These processes would be repeated a million times for accomplishing a data-intensive computing task; consequently, the communication bus connecting CPUs and memory inevitably becomes the energetic and speed bottleneck. Moreover, for achieving more powerful computing capability with low-power consumption, the transistor scaling and operating frequency increment is becoming the direction of technological development.

To achieve high computing capability, an extremely large number of transistors have been compressed in a single CPU. Furthermore, the power consumption is almost linear dependent proportionally with operation frequency [21]. This means that the power consumption and computing capability need to be balanced and cannot be achieved simultaneously with recent CMOS technology under the von Neumann architecture. On the contrary, scientists have noticed that the human brain has an excessive computing and energy efficiency [22]. With the idea and hypotheses to build a brain-like computing machine, the concept of neuromorphic computing was proposed by Dr. Mead [3]. The significance of the neuromorphic computing is not only for building a more-powerful computer, but also can potentially reveal the fundamental operating mechanism of the human brain. Another similar well-known concept is the artificial neural networks (ANNs), which is an attempt of simulating the neural network configuration of the brain, thereby to study the function of the brain [23, 24]. The main differences between neuromorphic computing and conventional ANNs are the former focuses more on the physical realization on the brain structure, while the latter studies the mathematical models of human brain structure. Neuromorphic computing is expected to offer an intelligent machine beyond the modern digital computer with capabilities of adaptive, distributive, cognitive computing, and perceptive computing. These capabilities fundamentally come from the unique architecture, computing/memory units, signal encoding scheme, and operating algorithms of the neuromorphic computing system.

To successfully implement a neuromorphic computing system, a comprehensive understanding of the differences between the human brain and von Neumann-based computer would be conducive to reverse engineering the brain, thus implementing the neuromorphic computing system. **Figure 3** illustrates the main difference between the human brain and the von Neumann architecture from the device to the algorithm levels. In a brain-like neuromorphic computing system, blocking devices (computing units and memory units) need to be replaced from traditional CPUs and SRAMs to artificial electronic neurons and synapses. This is the first step for mimicking the brain at a device level. Unlike computing units in the CPUs that perform the binary code–based computing, the data in electronic neurons and synapses need to be represented in a spike sequence format for generating the brain-like signals [22]. Then, these electronic neurons and synapse are interconnected with each other in a brain-like neural network configuration at the architecture level, which is demonstrated in **Figure 3**. Spiking signals would be used for communication in this architecture. This neural network-based architecture eliminates the long signal transferring distance between CPUs and the memory in von Neumann architecture since the computing can be performed by neurons with the data extracted from adjacent memories (synapses). Due to the unique non-von Neumann architecture and spiking encoding scheme of the neuromorphic computing system, the binary algebra is not suitable for this system anymore. In the field, neural network-based machine learning algorithms are widely considered as the ideal candidate for running neuromorphic computing system.

Although fundamental functions of the brain are still under investigation, two main elements: neuron and synapse are well studied at the cellular level. The structure of a neuron is shown in **Figure 4**. There are four main parts of each neuron, whose functionalities are summarized as:

- Dendrite: the organ that receives spiking signals from other neurons,
- Soma (neuron body): generates/sends spiking signals to the axon under the condition of the integration of received spiking signal levels, which exceed a specific threshold voltage;
- Axon: propagates spiking signals to other neurons,
- Synapse: a space between the axon of the presynapse neuron and dendrite of the postsynapse neuron. It is widely considered as a memory organ in the brain by storing the memory information in its connectivity strength.

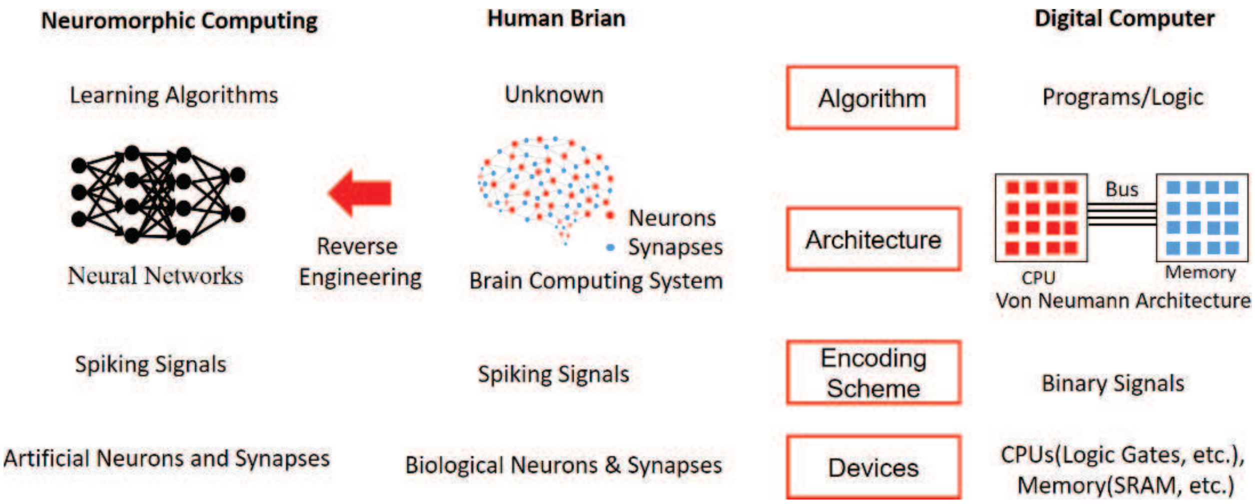

Figure 3. Comparison between brain computing architecture, von Neumann computing architecture, and neuromorphic computing architecture.

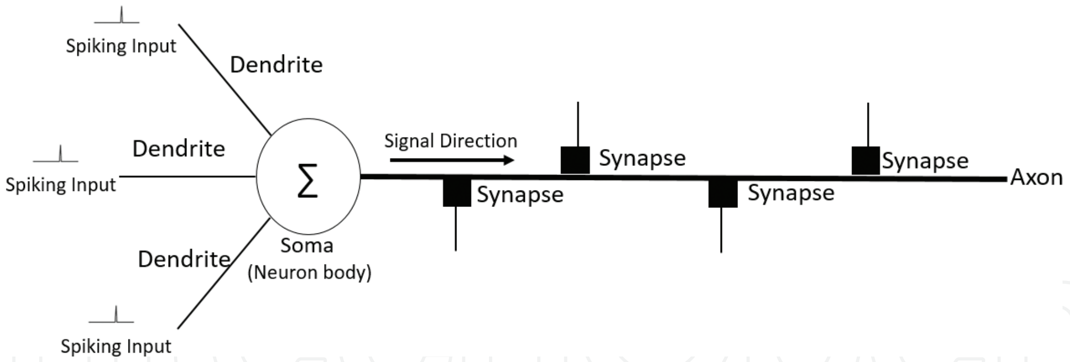

Figure 4. Neuron structure.

Unlike the rigid connection configuration of computing units and memory in the von Neumann architecture (**Figure 2**), neurons and synapses can be connected to each other in different topologies. **Figure 5** depicts three mainstream neuromorphic computing architectures named as the distributed neuromorphic computing architecture, cluster neuromorphic computing architecture, and associative neuromorphic computing architecture [13].

Firstly, the distributed neuromorphic computing architecture (DNCA) decomposes centralized computing units and memory units in a distributed brain-like network structure. In this architecture, neurons and synapses are located close to each other to minimize the signal propagation distance through communicating only with the adjacent electronic synapse (memory data).

Secondly, in the human brain, different types of sensory signals (for example somatic, tactile, auditory, visionary, olfactory, and gustatory signals) are routed and processed in different regions of the brain [22]. The cluster neuromorphic computing architecture (CNCA) is proposed to realize this signal processing methodology of the human brain. In this architecture, the proposed DNCA is divided into multiple regions, which are intrinsically responsible for processing signals captured by different types of sensory devices independently. This signal processing technique enables the CNCA to process multiple massive signals parallel in various regions with distributedly located neurons and synapses, thereby, realizing a parallel computing capability inherent similar to the human brain.

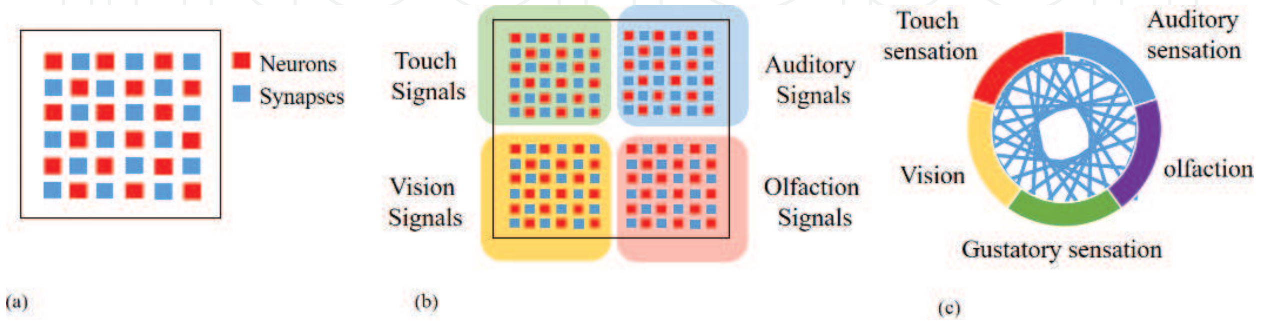

Figure 5. The neuromorphic computing architectures: (a) distributive neuromorphic computing architecture, (b) cluster neuromorphic computing architecture, and (c) associative neuromorphic computing architecture [13].

Thirdly, the human brain has a powerful unsupervised learning ability, which enables us to learn from our experiences. A well-known learning mechanism named associate memory is to associate different types of signals captured by various sensing organs together [22] so that it correlates these signals. Based on the CNCA, a novel architecture, we defined it with the name of associative neuromorphic computing architecture (ANCA), is proposed. **Figure 5(c)** illustrates this architecture. In this architecture, original signals captured from surrounding environments are processed in different regions. After that, the abstracted information would couple to each other to construct an associative natural network. The simplified ANCA with two neurons and one synapse has been investigated [25].

3. Neuron design

3.1. Neuron models

In the field of neuroscience, the research on the investigation of biological neurons has been continued in the past decade [26–31]. As discussed in Section 2, a neuron consists of four major elements, namely, dendrites, soma, axon, and synapse. Within the nervous system, signals are collected and transmitted to the soma by dendrites. The soma serves as the central processing unit where the nonlinear transformation carries out. When the input signal exceeds the threshold level, an output signal is generated, or so-called the firing process. The output signal is then transmitted along the axon, and to other neurons through the synapse. In a biological neuron, signals are in form of a nerve impulse, namely, action potential or spike [32].

When the signal, also known as the stimulus, from dendrites does not reach the critical threshold level, the membrane potential will leak out; otherwise, an action potential is generated. After the firing process takes place, the neuron will go through a refractory period, where the neuron is less likely to fire, and eventually reset to its initial state. This process is known as the firing and resting of a biological neuron, as illustrated in **Figure 6** [31]. Several well-known and representative neuron models are investigated, which include the integrate-and-fire (IF) model [26], Fitzhugh-Nagumo (FF) model [28], Hodgkin-Huxley (HH) model [33], and leaky integrate-and-fire (LIF) model [29]. The simplified electronic circuit representation of these neuron models is demonstrated in **Figure 7**.

3.2. Hodgkin-Huxley (HH) and Fitzhugh-Nagumo (FN) neuron model

Compared to the data that are extracted from the IF neuron, the HH neuron is found to be biologically meaningful and realistic [34]. The primary goal of the HH neuron is to mimic the electrochemical information transmission of a biological neuron [27]. **Figure 7(c)** demonstrates the simplified electronic circuit model of the HH neuron. The dynamic of the firing potential is described by a fourth-order nonlinear differential equation, which could be simplified as

$$C_m \cdot \frac{dV_m}{dt} = I_{ex} - g_i(h, m^3, n^4) \cdot \sum I_i(E_i, V_m), \tag{1}$$

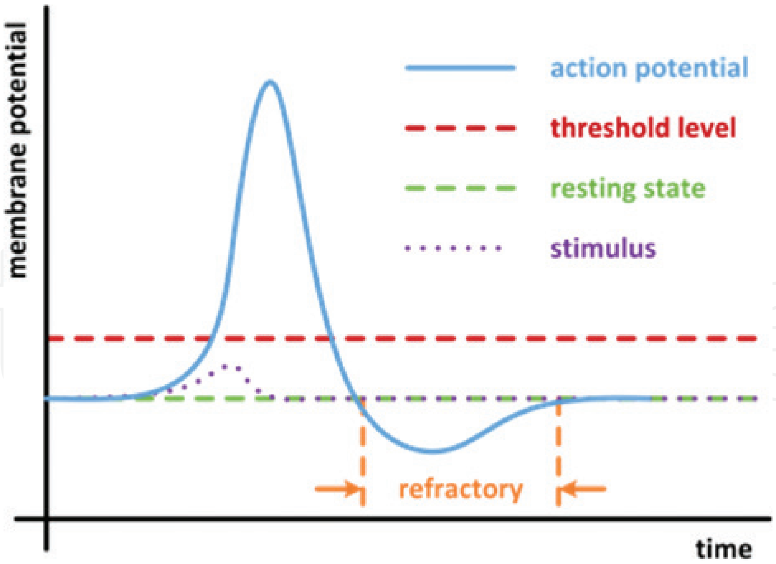

Figure 6. Action potential of a biological neuron.

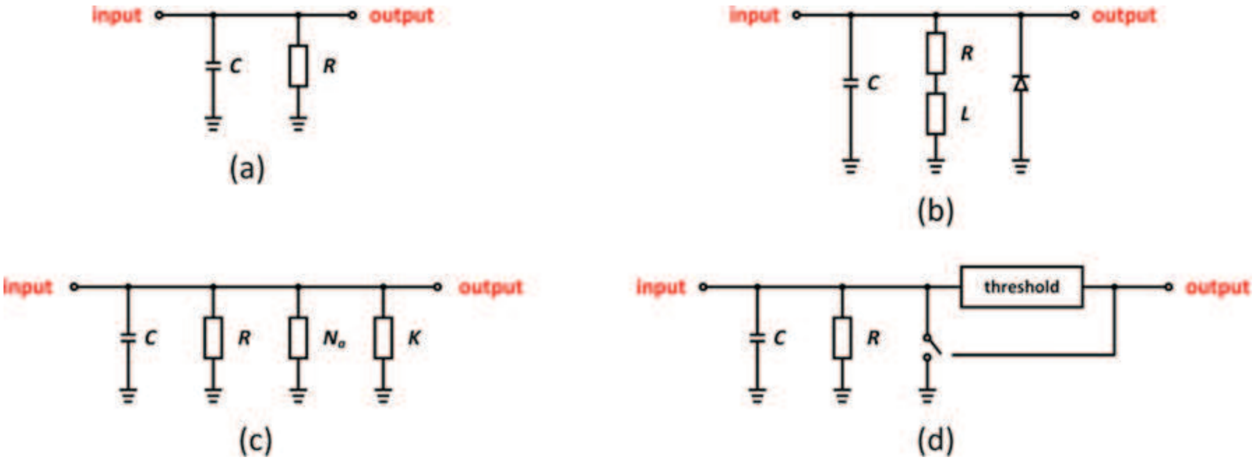

Figure 7. Simplified neuron models of (a) integrate-and-fire, (b) Fitzhugh-Nagumo, (c) Hodgkin-Huxley, and (d) leaky integrate-and-fire.

where g_i is the conductance parameter for different ion channels (sodium Na, potassium K, etc.), and $I_i(E_i, V_m)$ is the ion current with controlling variable as a function of time [33]. Although the HH neuron closely mimics the biological behavior of neurons, due to its design complexity, its electronic circuit model is not widely used in the hardware implementation, whereas the FN neuron is considered as the simplification of the HH neuron, as shown in **Figure 7(b)**. Its mathematical expression could be written as

$$\frac{dV_m}{dt} = V_m - \frac{V_m^3}{3} - w + I_{ex}, \tag{2}$$

where w is the linear recovery variable. Although the FN neuron reduces the four-dimensional set of the equations down to a two-dimensional one, the hardware implementation of the FN neuron is still excessive challenging due to its high circuit design complexity inherent from its highly nonlinear behavior.

3.3. Leaky integrate-and-fire (LIF) neuron model

The LIF neuron model, as illustrated in **Figure 7(d)**, is constructed based on the traditional IF neuron. Its leakage property mimics the diffusion of ions that occur through the membrane when the equilibrium is not reached in the cell. The dynamic of the firing potential could be expressed as:

$$C_m \cdot \frac{dV_m}{dt} + I_{leak} = I_{ex}, \tag{3}$$

where I_{leak} is the leakage current. Similar to the traditional IF neuron, the membrane potential is initially charged up by the excitation current. An action potential is generated once the membrane potential exceeds the threshold level; otherwise, all charges will be leaked out. After the firing process takes place, the membrane capacitor in the LIF neuron will be fully discharged to the resetting state. Hence, the LIF neuron processes both firing and resting properties, which has an adequate resemblance to the biological neuron and relatively easier to implement using analog electronic circuits.

Compared to other neuron models, the LIF neuron plays a major role in the neuron design due to its compact structure, robust performance, and adequate resemblance to the biological behavior of neurons. The simplified analog electronic circuit model of the LIF neuron is demonstrated in **Figure 8**.

In the analog electronic circuit model of the LIF neuron, there are several key parameters that need to be carefully designed; for instance, the excitation current I_{ex}, the membrane capacitor C_m, the threshold level V_{th}, and the leakage current I_{leak}. In Eq. (4), the membrane potential is controlled by the excitation current and the leakage current, or vice versa. A simple resistor model is adapted to represent such relation; thus, Eq. (3) could be rewritten as

$$I_{ex} = \frac{V_m}{R_{leak}} + C_m \cdot \frac{dV_m}{dt}, \tag{4}$$

where R_{leak} defines the weighted resistance of the leakage current. By solving Eq. (4), the expression of the membrane potential could be determined as

$$V_m = I_{ex} \cdot R_{leak} - e^{\frac{t}{R_{leak} \cdot C_m}}. \tag{5}$$

3.4. Signal intensity encoding neuron

In order to model the input intensity-dependent firing characteristic of neurons [22, 35], the signal intensity encoding neuron (SIEN) is designed, as depicted in **Figure 9** [36].

In this design, the input current is transferred into a voltage signal by a transimpedance amplifier (TIA), such that the oscillating frequency of a current-starved-voltage controlled

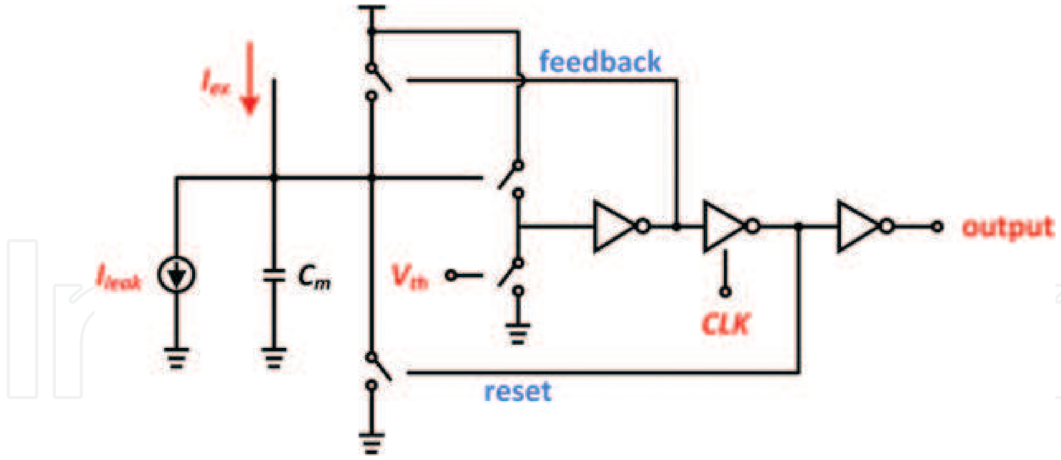

Figure 8. Simplified analog electronic circuit model of the LIF neuron.

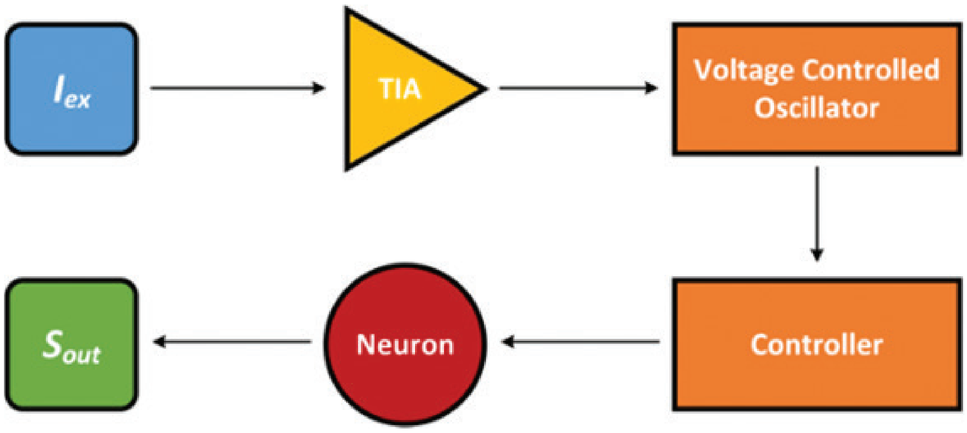

Figure 9. The diagram of the SIEN.

oscillator (VCO) can be regulated. The oscillating rate of the VCO is highly dependent upon integrated input stimulus signals. The final stage of the SIEN is formed by the parallel structure of a resistor and a capacitor to model charging and discharging behaviors of biological neurons, as depicted in **Figure 10**, whereas simulation results of the spiking signal are plotted

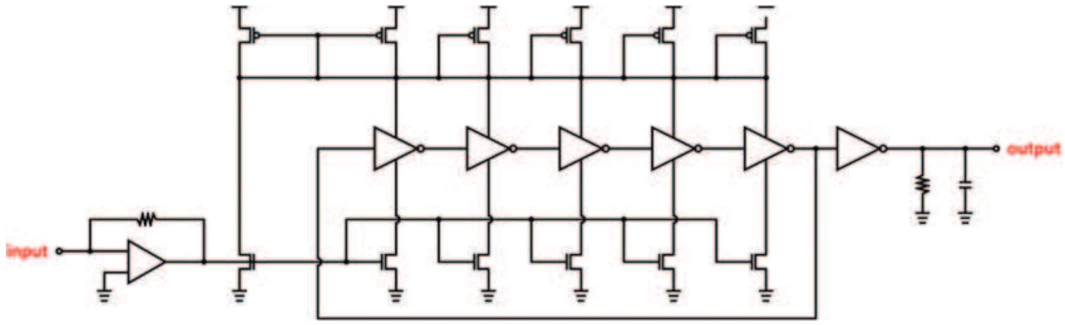

Figure 10. Simplified design scheme of the SIEN.

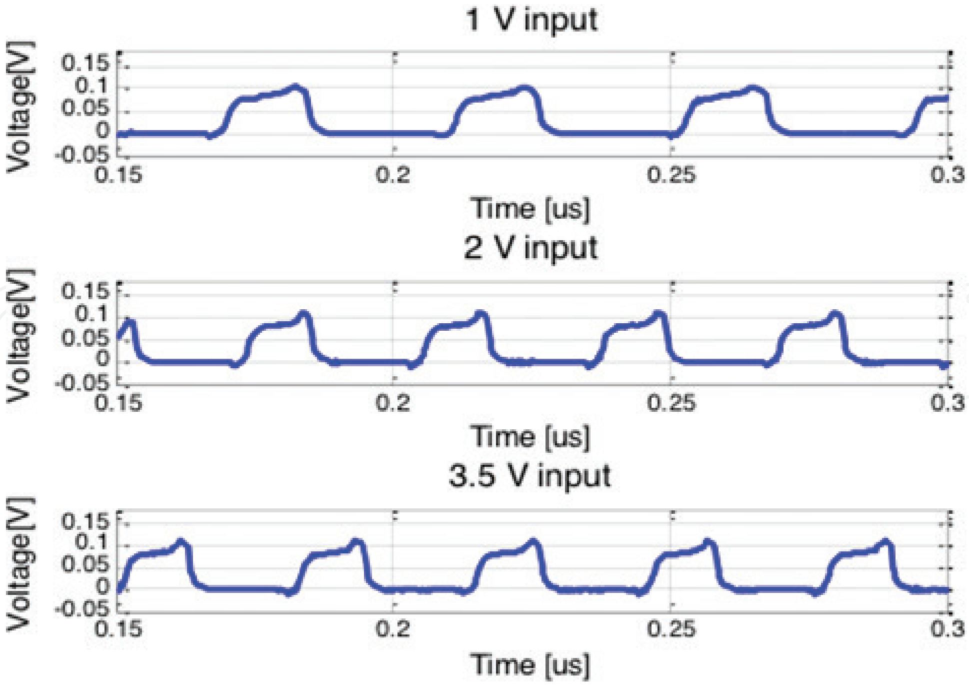

Figure 11. Spiking signals with respect to various stimulus voltage levels.

in **Figure 11**. In **Figure 11**, with higher input signal amplitudes, the firing rate increases correspondingly, which simulates the input intensity-dependent firing characteristic of the neurons in real biological systems.

4. Memristor as synapse

In the human brain, a synapse is defined as the structure connecting two neurons as shown in **Figure 12**. When a presynaptic action potential (spiking signal) approaches to the synapse, the chemical neurotransmitter molecules would be released to the synapse. The neurotransmitter would be diffused across from the presynaptic neuron cell to the postsynaptic neuron cell within the synapse. When the neurotransmitter arrived at the terminal of the postsynaptic cell, a spiking signal would be stimulated. The magnitude of the stimulated spiking signal

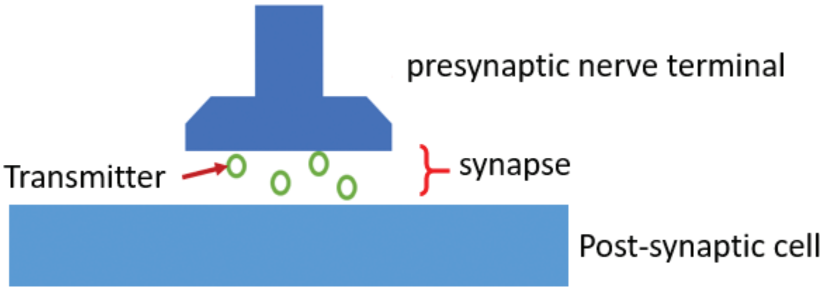

Figure 12. The structure of the synapse [22].

at the postsynaptic cell is highly dependent on the amount of the neurotransmitter received. A larger amount of neurotransmitter molecules stimulate a larger magnitude spiking signal, vice versa. In general, the large magnitude of spiking signal at the terminal of presynaptic neurons would stimulate more neurotransmitters. However, with the repeated stimulus in a short time (~hundreds of millisecond), the neurotransmitters released to the synapse from presynaptic neurons reduce gradually, which results in stimulating a smaller magnitude spiking signal in the postsynaptic neuron.

This phenomenon was investigated by Dr. Kandel's research on Aplysia [22]. In experiments depicted in **Figure 13**, the stimulus was repeatedly applied to the Aplysia's sensory neurons. When the constant stimulus was repeatedly applied to the sensory neuron multiple times (1, 2, 5, 10, 15), the magnitude of spiking signal stimulated in the response neuron ($L7_G$) decreases accordingly [22]. This indicates that the previous stimulus captured by the sensor neuron is somehow stored in the neural network system through modifying the connectivity strength between neurons. In Dr. Kandel's experiments, the neural network is relatively small that is only constructed by two neurons. The connectivity strength of the synapse is defined as the weight. The weight value can be modified in two directions (strengthen or weaken) by both excitatory and inhibitory stimuli. This feature is called as the plasticity of a synapse.

In order to physically realize the biological plasticity of a synapse, several features need to be satisfied. Firstly, the device should have only two terminals that are used for connecting the presynaptic and postsynaptic neurons, respectively. Secondly, the device should have a signal attenuation capability to mimic the plasticity of a synapse, and this capability should be reversible. All these features make the nanoscale two-terminal device memristor, also named as resistive RAM (RRAM), to be an ideal candidate for the electronic synapse implementation. The resistance of the memristor is reversibly programmable with the applied voltage pulse stimulus on its two terminals. When the voltage stimulus is applied on its two terminals, its resistance would be gradually changed between its low-resistance state (LRS) and high-resistance state (HRS). Typically, the memristor is constructed by the metal-insulator-metal (MIM) configuration as illustrated in **Figure 14(a)**. The decrease of resistance of the memristor is due to the formation of the conductive filament in the insulator layer. Transmission electron microscopy (TEM) photos of conductive filaments are demonstrated in **Figure 14(b)**. This breakdown phenomenon of the insulator can be recovered by applying

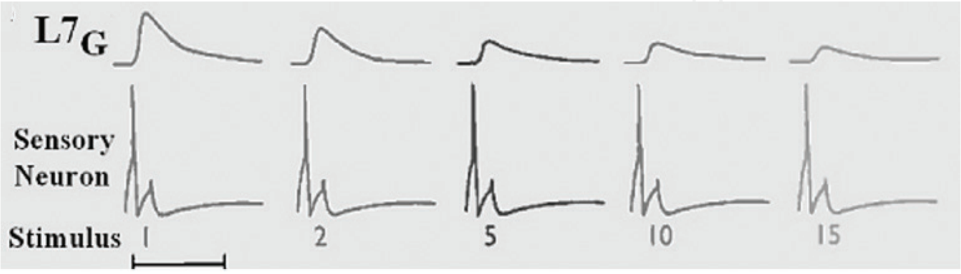

Figure 13. A sample of five identical action potential numbers 1, 2, 5, 10, and 15 along with the corresponding motor response signals of diminishing strength recorded at the motor neuron (identified by $L7_G$) (top) [37].

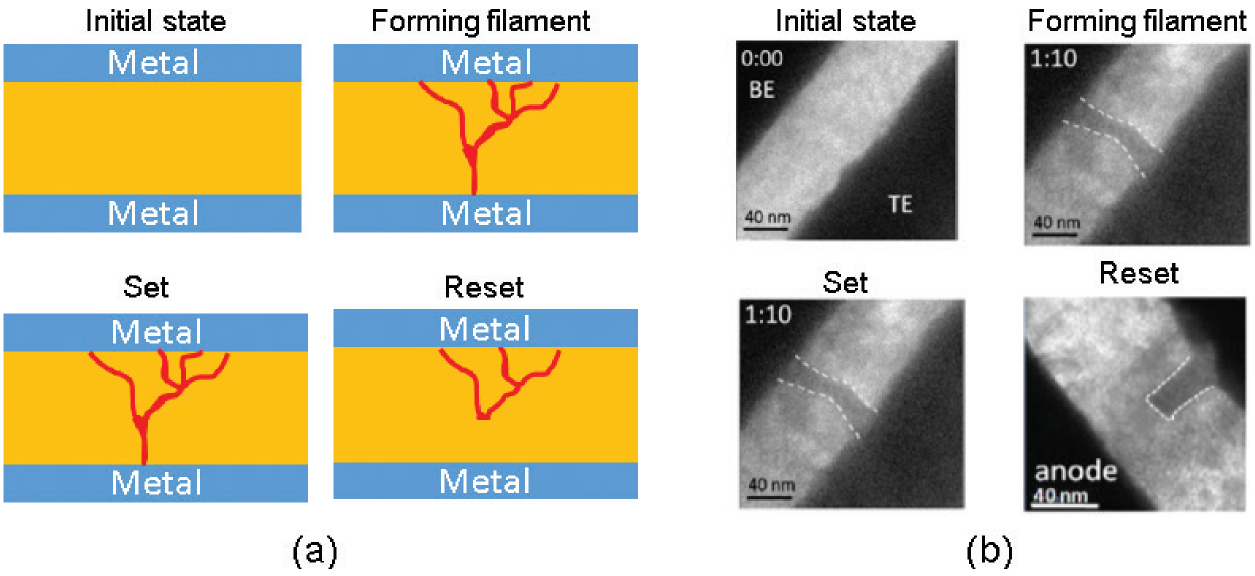

Figure 14. Illustration of the switching mechanism of a memristor: (a) switching process and (b) TEM images of the dynamic evolution of conductive filaments [38].

a reversed stimulus at the terminals, which consequently resets the memristor from its LRS to HRS. The physical mechanism of this reset behavior is the deconstruction of the conductive filament as illustrated in **Figure 14(b)**.

In general, the MIM structure of the memristor is fabricated massively in a 2D crossbar structure as depicted in **Figure 15**. In this structure, memristors are sandwiched between two layers of nanowires. The area of a single cell is 4F^2, where the F is the minimum lithographic feature size dictated by technology node.

In order to further enhance the device density, the 2D crossbar structure of the memristor can be extended vertically into 3D space. There are two types of 3D RRAM (memristor) structures that can be used as 3D synaptic arrays: horizontal RRAM (H-RRAM) and vertical RRAM (V-RRAM), which are shown, respectively, in **Figure 16**.

In both structures, the area of the device size is $4F^2/n$, where n is the number of the stacked layers. The number of critical lithography masks for H-RRAM structure increases linearly with

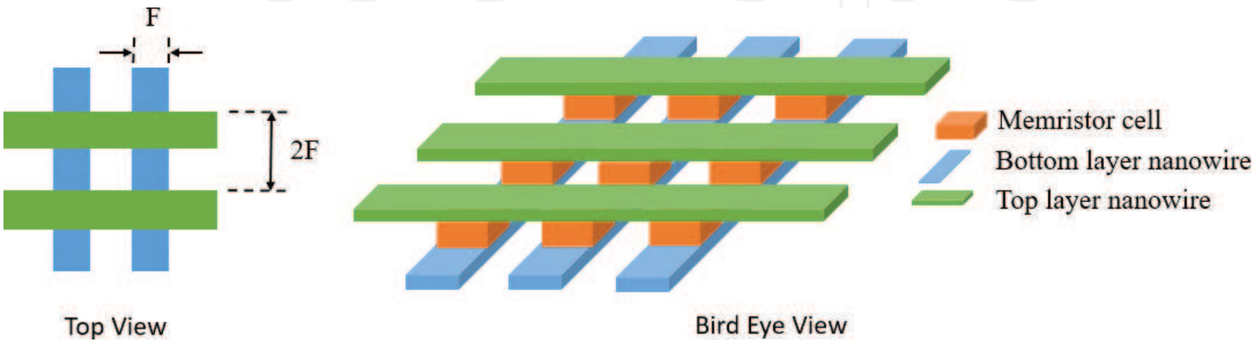

Figure 15. Two-dimensional crossbar structure of the memristor.

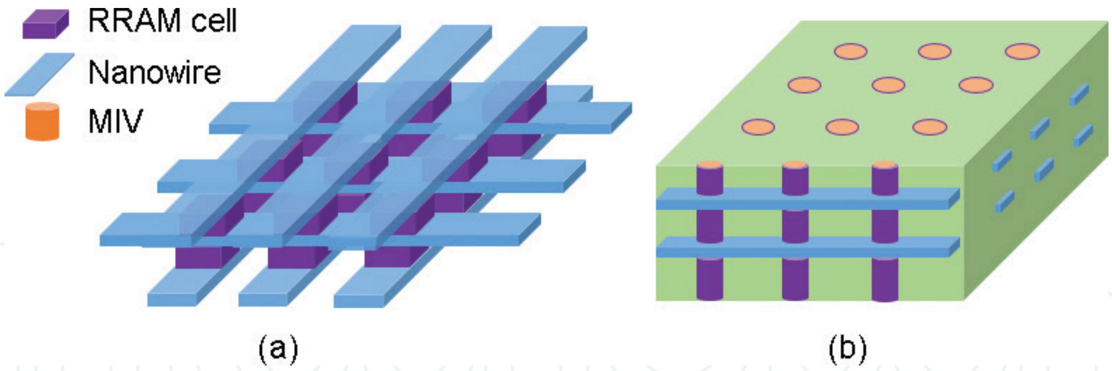

Figure 16. 3D RRAM integration structure: (a) horizontal 3D structure and (b) vertical 3D structure.

increasing number of the stacked layers, while the number of masks for V-RRAM is relatively independent of the stacking number. With increasing number of the stacked layers, V-RRAM becomes, even more, cost-effective [39, 40].

5. Memristive three-dimensional neuromorphic computing system

The recent fabricated neuromorphic chips implement neurons and synapses using traditional 2D CMOS and memory technology. In 2D placement methodology, a longer signal delivery distance is generally expected due to the routing density increment linearly with the number of connections, which inevitably increases die area, power consumption, etc. [41].

To address these limitations of the state-of-the-art neuromorphic chip designs, a novel 3D neuromorphic architecture is proposed to combine 3D-integrated circuit (3D-IC) technology with the memristor as the electronic synapse. Applying 3D integration technology to neuromorphic chips permits vertical routing paths of reduced nanoscale dimension, subsequently diminishing critical path lengths. It also decreases power consumption and shrinks die areas with high-complexity, high-connectivity, and massively parallel signal processing capability.

The benefits of applying 3D integration technology to neuromorphic chips design can be summarized as follows:

1. address the 2D neuron routing congestion problem, thereby increasing interconnectivity and scalability of the NC network and reducing the critical-path lengths [42],
2. allow numerous 3D interconnections between hardware layers that offer high device interconnection density, low-power density, and broad channel bandwidth using fast and energy-efficient links;
3. provide a high-complexity, high-connectivity, and massively parallel-processing circuital system that can accommodate highly demanding computational tasks.

The diagram structure of the proposed 3D neuromorphic computing (3D-NC) architecture is shown in **Figure 17(c)**. The multiple layers of the neural network can be implemented through this structure. **Figure 17(a)** illustrates multiple layers of the neural network structure, in which the decomposed two layers are marked in a red box. These two layers of the neural network can be implemented through 3D integration technology, which fabricates the layer of memristor in the middle between two neuron layers as depicted in **Figure 17(b)**. Besides, with the similar structure of **Figure 17(b)**, a large scale of neural networks can be implemented by extending the 3D structure of two layers neural network repeatedly in a horizontal direction, which is demonstrated in **Figure 17(c)**.

In these structures, the electronic synaptic array implemented with memristors is not in a traditional crossbar structure (**Figure 18(a)**), which suffers the sneaking path issue. The sneaking path is an undesired current path from the adjacent memristor cells marked as the white arrows in **Figure 18(a)**. In order to eliminate this issue, the horizontal nanowires, which are used for reading/writing access, are physically disconnected in the design. Meanwhile, reading/writing access ports are located on the upper and bottom layers. Without electrical connections between adjacent memristor cells, the sneaking path issue can be fundamentally addressed.

Two 3D integration technologies have the potential for implementing the 3D-NC architecture in **Figure 17(c)**, which are TSV (through-silicon via)-based and monolithic-based 3D integration technologies. The 3D integration technology with TSVs as vertical electrical connections has been studied for many years [19]. For TSV-based 3D integration technology, transistors are initially fabricated at separated wafers by traditional CMOS technologies. After that, two wafers are bonded together. In general, the capacitance between TSVs is large, which can cause capacitive coupling issue in a high-speed circuit. However, they can be used for implementing the capacitance in neuron models, resulting in further reduction of the chip design area [43–45]. However, there are several technical challenges for the TSV-based 3D integration technology. Firstly, wafers need to be thinned to make the metal contact from TSVs for

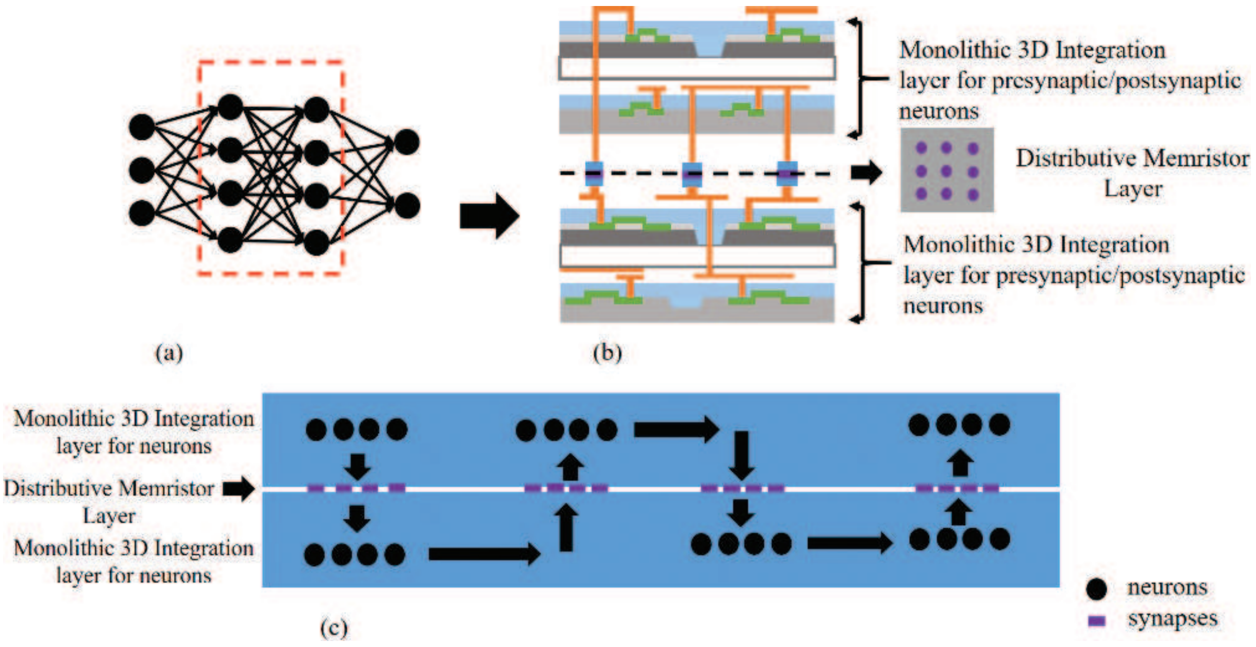
Monolithic 3D Integration layer for presynaptic/postsynaptic neurons
Distributive Memristor Layer
Monolithic 3D Integration layer for presynaptic/postsynaptic neurons
(a)
(b)
Monolithic 3D Integration layer for neurons
Distributive Memristor Layer
Monolithic 3D Integration layer for neurons
neurons
synapses
(c)

Figure 17. 3D neuromorphic computing architecture (a) Deep neural network, (b) 3D structure of two layers of neural network and (c) 3D structure of multiple layers of neural network.

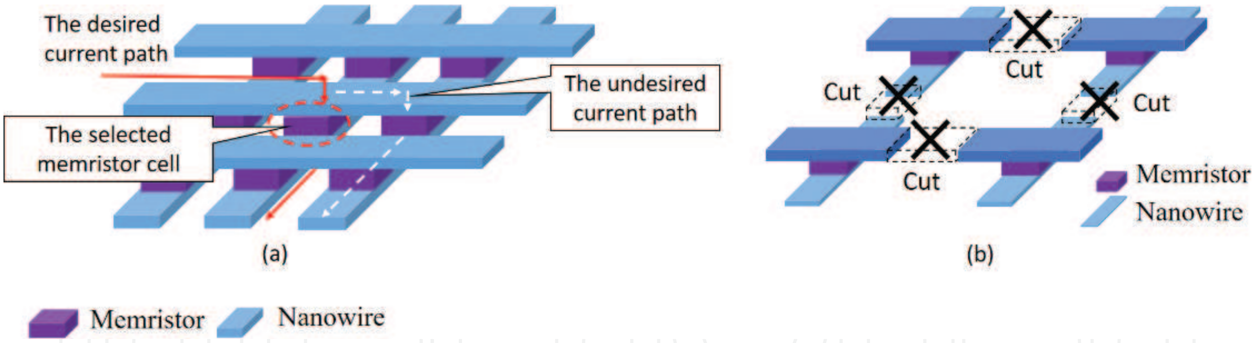

Figure 18. (a) The traditional crossbar structure of memristors and (b) disconnecting the horizontal connecting nanowire.

3D device	FinFET	Epi-like Si NWFET	Epi-like Si UTB	SOI-Si UTB	Poly-Si/Ge FinFET	IGZO OSFET
Thermal budget ˚C	<400	<400	<400	<650	<400	<500
I_on/I_off	$>10^7$	$>5 \times 10^5$	$>5 \times 10^5$	$>10^7$	$>10^7$	$>10^{21}$

Table 1. The emerging transistors with low-fabrication temperature [46].

the bonding process. In these thinning processes, a lot of charges would be accumulated. These charges potentially cause electrostatic discharge (ESD) issue that can damage chips in bonding processes. Secondly, bonding the microscale TSVs needs extra effort to align them precisely. To overcome these challenges, another more aggressive 3D integration technology is proposed, which is called monolithic 3D integration. Unlike the TSV-based 3D technology, which uses a separate fabricate processes, the monolithic 3D technology integrates different layers of devices at a single wafer with nanoscale intertier vias serving as vertical connections. Due to the monolithic fabrication procedure, this 3D integration technology fundamentally eliminates the thinning and bonding processes. On the contrary, the main challenge for the monolithic 3D integration technology is the low-temperature fabrication constraint for upper layers, since the high fabrication temperature in upper layers would damage the lower layer transistors previously fabricated. This low-temperature requirement restricts the traditional CMOS transistor (fabricates at more than 1000 ℃) that does not fit the requirements for the upper layer circuitry implementation. Fortunately, several low-temperature transistors are the potential candidate to fit this requirement, such as FinFETs [46], carbon nanotube FETs [47, 48], etc. **Table 1** summarizes state-of-the-art transistors that are fabricated at low temperature and potentially can be employed in the monolithic 3D integration technology [46]. With these emerging technologies, the monolithic 3D-NC with memristors as electronic synapses is becoming the most promising next-generation non-von Neumann computing platform.

6. Conclusion

The conventional concept of the neuromorphic computing is to physically rebuild brain-like neural networks through very-large-scale integration (VLSI) [3]. In this chapter, we introduce a possibility to use an emerging device named memristor as an electronic synapse to construct a memristive neural network of the neuromorphic computing system, consequently, achieving a much smaller design area and power consumption. In this chapter, we also comprehensively analyze functions of the biological synapse in cellular level and further introduce the reasons that memristor can be considered as an electronic synapse. In architecture level of neuromorphic computing, we introduce three novel architectures that are fundamentally different from the traditional von Neumann architecture by locating the computing units (neurons) and memory units (synapse) distributedly. The realization of these three neuromorphic computing architectures potentially is a roadmap for implementing a power-efficient artificial intelligent system with self-learning capability.

Furthermore, the memristive neural network is generally implemented in the two-dimensional design method. In this chapter, we introduce and discuss a novel hardware implementation trend that combines memristor and 3D-IC integration technology; such technology has the capabilities to reduce the system power consumption, provide the high-connectivity, resolve the routing congestion issues, and offer the massively parallel data processing capability. Moreover, the design methodology of applying the capacitance formed by the through-silicon vias (TSVs) to generate a membrane potential in a 3D neuromorphic computing system is discussed in this chapter.

Moreover, there are several challenges that hinder the employment of the memristors as the electronic synapse, e.g., the reliability, variability, endurance, etc. Additionally, fabrication techniques of lower temperature transistors (FinFET, carbon nanotube FETs, etc.), which can be integrated monolithically on the top layers, demand further research effort to demonstrate the memristive 3D neuromorphic computing system discussed in this chapter. The proposed novel neuromorphic computing architectures (DNCA, CNCA, and ANCA) are considered potentially to be the roadmap for achieving a self-learning artificial intelligence that can directly learn from the surrounding environment and be adaptive to it. However, mathematical foundations of these architecture concepts are still unclear and missing, which need further investigations in future.

Author details

Hongyu An*, Kangjun Bai and Yang Yi

*Address all correspondence to: hongyu51@vt.edu

The Bradley Department of Electrical and Computer Engineering, Virginia Polytechnic
Institute and State University, Blacksburg, Virginia, USA

References

[1] Kish LB. End of Moore's law: Thermal (noise) death of integration in micro and nano electronics. Physics Letters A. 2002;**305**:144-149

[2] Yu S, Wu Y, Jeyasingh R, Kuzum D, Wong H-SP. An electronic synapse device based on metal oxide resistive switching memory for neuromorphic computation. IEEE Transactions on Electron Devices. 2011;**58**:2729-2737

[3] Mead C. Neuromorphic electronic systems. Proceedings of the IEEE. 1990;**78**:1629-1636

[4] Ghani A. Neuro-inspired speech recognition based on reservoir computing. Advances in Speech Recognition. InTech; 2010;**2**:7-36

[5] Overton G. Photonic Reservoir Computing–A New Tool for Speech Recognition. https://www.laserfocusworld.com/articles/2014/09/photonic-reservoir-computing-a-new-tool-for-speech-recognition.html

[6] Alalshekmubarak A, Smith LS. On improving the classification capability of reservoir computing for Arabic speech recognition. In: International Conference on Artificial Neural Networks; 2014. pp. 225-232

[7] Verstraeten D, Schrauwen B, Stroobandt D. Reservoir computing with stochastic bitstream neurons. In: Proceedings of the 16th annual Prorisc Workshop; 2005. pp. 454-459

[8] Jin Y, Zhao Q, Yin H, Yue H. Handwritten numeral recognition utilizing reservoir computing subject to optoelectronic feedback. In: Natural Computation (ICNC), 2015 11th International Conference on; 2015. pp. 1165-1169

[9] Hinaut X, Dominey PF. On-line processing of grammatical structure using reservoir computing. In: International Conference on Artificial Neural Networks; 2012. pp. 596-603

[10] Goudarzi A, Lakin MR, Stefanovic D. Reservoir computing approach to robust computation using unreliable nanoscale networks. In: International Conference on Unconventional Computation and Natural Computation; 2014. pp. 164-176

[11] Jaeger H. Short Term Memory in Echo State Networks vol. 5. GMD-Forschungszentrum Informationstechnik. Germany: Schloss Birlinghoven 53757 Sankt Augustin; 2001

[12] Schrauwen B, Stroobandt D. Using reservoir computing in a decomposition approach for time series prediction. In: ESTSP 2008 European Symposium on Time Series Prediction; 2008. pp. 149-158

[13] An H, Zhou Z, Yi Y. Opportunities and challenges on nanoscale 3D neuromorphic computing system. In: Electromagnetic Compatibility & Signal/Power Integrity (EMCSI), 2017 IEEE International Symposium on; 2017. pp. 416-421

[14] Qiao N, Mostafa H, Corradi F, Osswald M, Stefanini F, Sumislawska D, et al. A re-configurable on-line learning spiking neuromorphic processor comprising 256 neurons and 128K synapses. Frontiers in Neuroscience. April 2015;**9**:141

[15] Benjamin B, Gao P, McQuinn E, Choudhary S, Chandrasekaran AR, Bussat JM, et al. Neurogrid: A mixed-analog-digital multichip system for large-scale neural simulations. Proceedings of the IEEE. May 2014;**102**:699-716

[16] Painkras E, Plana LA, Garside J, Temple S, Galluppi F, Patterson C, et al. SpiNNaker: A 1-W 18-core system-on-chip for massively-parallel neural network simulation. IEEE Journal of Solid-State Circuits. 2013;**48**:1943-1953

[17] Furber SB, Lester DR, Plana LA, Garside JD, Painkras E, Temple S, et al. Overview of the SpiNNaker system architecture. IEEE Transactions on Computers. 2013;**62**:2454-2467

[18] An H, Zhou Z, Yi Y. 3D memristor-based adjustable deep recurrent neural network with programmable attention mechanism. In: Proceedings of Neuromorphic Computing Symposium; 17-19 July 2017. pp. 1-6

[19] Ehsan MA, Zhou Z, Yi Y. Modeling and optimization of TSV for crosstalk mitigation in 3D neuromorphic system. In: Electromagnetic Compatibility (EMC), 2016 IEEE International Symposium on; 2016. pp. 621-626

[20] Koyanagi M, Nakagawa Y, Lee K-W, Nakamura T, Yamada Y, Inamura K, et al. Neuromorphic vision chip fabricated using three-dimensional integration technology. In: Solid-State Circuits Conference, 2001. Digest of Technical Papers. ISSCC. 2001 IEEE International; 2001. pp. 270-271

[21] Merolla PA, Arthur JV, Alvarez-Icaza R, Cassidy AS, Sawada J, Akopyan F, et al. A million spiking-neuron integrated circuit with a scalable communication network and interface. Science. Aug 2014;**345**:668-673

[22] Kandel ER, Schwartz JH, Jessell TM, Siegelbaum SA, Hudspeth A. Principles of Neural Science Vol. 4. New York: McGraw-Hill; 2000

[23] Turing AM. On computable numbers, with an application to the Entscheidungsproblem. Proceedings of the London Mathematical Society. 1937;**2**:230-265

[24] McCulloch WS, Pitts W. A logical calculus of the ideas immanent in nervous activity. The Bulletin of Mathematical Biophysics. 1943;**5**:115-133

[25] An H, Zhou Z, Yi Y. Memristor-based 3D neuromorphic computing system and its application to associative memory learning. In: 2017 IEEE 17th International Conference on Nanotechnology (IEEE-NANO); 2017. pp. 555-560

[26] Abbott LF. Lapicque's introduction of the integrate-and-fire model neuron (1907). Brain Research Bulletin. 1999;**50**:303-304

[27] Hodgkin AL, Huxley AF. A quantitative description of membrane current and its application to conduction and excitation in nerve. The Journal of Physiology. 1952;**117**:500

[28] FitzHugh R. Impulses and physiological states in theoretical models of nerve membrane. Biophysical Journal. 1961;**1**:445-466

[29] Liu Y-H, Wang X-J. Spike-frequency adaptation of a generalized leaky integrate-and-fire model neuron. Journal of Computational Neuroscience. 2001;**10**:25-45

[30] Zhao C, Danesh W, Wysocki BT, Yi Y. Neuromorphic encoding system design with chaos based CMOS analog neuron. In: Computational Intelligence for Security and Defense Applications (CISDA), 2015 IEEE Symposium on; 2015. pp. 1-6

[31] Zhao C, Yi Y, Li J, Fu X, Liu L. Interspike-interval-based analog spike-time-dependent encoder for neuromorphic processors. IEEE Transactions on Very Large Scale Integration (VLSI) Systems. 2017;**25**:2193-2205

[32] Gerstner W, Kistler WM. Spiking Neuron Models: Single Neurons, Populations, Plasticity. New York, USA: Cambridge University Press; 2002

[33] Merolla P, Arthur J, Akopyan F, Imam N, Manohar R, Modha DS. A digital neurosynaptic core using embedded crossbar memory with 45pJ per spike in 45nm. In: Custom Integrated Circuits Conference (CICC), 2011 IEEE; 2011. pp. 1-4

[34] Paugam-Moisy H. Spiking Neuron Networks a Survey. Switzerland: IDIAP; 2006

[35] Zhang Y, Igwe OJ. Exogenous oxidants activate nuclear factor kappa B through toll-like receptor 4 stimulation to maintain inflammatory phenotype in macrophage. Biochemical Pharmacology. Jan 2018;**147**:104-118

[36] An H, Ehsan MA, Zhou Z, Shen F, Yi Y. Monolithic 3D neuromorphic computing system with hybrid CMOS and memristor-based synapses and neurons. Integration, the VLSI Journal. 1 Nov, 2017

[37] Chua L. Memristor, Hodgkin-Huxley, and edge of chaos. In: Memristor Networks. New York: Springer; 2014. pp. 67-94

[38] Chen JY, Hsin CL, Huang CW, Chiu CH, Huang YT, Lin SJ, et al. Dynamic evolution of conducting nanofilament in resistive switching memories. Nano Letters. Aug 2013; **13**:3671-3677

[39] Xu C, Niu D, Yu S, Xie Y. Modeling and design analysis of 3D vertical resistive memory—A low cost cross-point architecture. In: 2014 19th Asia and South Pacific Design Automation Conference (ASP-DAC) 2014. pp. 825-830

[40] An H, Ehsan MA, Zhou Z, Yi Y. Electrical modeling and analysis of 3D synaptic array using vertical RRAM structure. In: Quality Electronic Design (ISQED), 2017 18th International Symposium on; 2017. pp. 1-6

[41] Akopyan F, Sawada J, Cassidy A, Alvarez-Icaza R, Arthur J, Merolla P, et al. True north: Design and tool flow of a 65 mW 1 million neuron programmable neurosynaptic chip. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems. Oct 2015;**34**:1537-1557

[42] An H, Ehsan MA, Zhou Z, Yi Y. Electrical modeling and analysis of 3D neuromorphic IC with monolithic Inter-tier vias. In: Electrical Performance of Electronic Packaging and Systems (EPEPS), 2016 IEEE 25th Conference on; 2016. pp. 87-90

[43] Ehsan MA, An H, Zhou Z, Yi Y. Adaptation of enhanced TSV capacitance as membrane property in 3D brain-inspired computing system. In: Proceedings of the 54th Annual Design Automation Conference; 2017. p. 86

[44] Ehsan MA, Zhou Z, Yi Y. Hybrid three-dimensional integrated circuits: A viable solution for high efficiency neuromorphic computing. In: VLSI Design, Automation and Test (VLSI-DAT), 2017 International Symposium on; 2017. pp. 1-2

[45] Yi Y, Zhou Y. Differential through-silicon-vias modeling and design optimization to benefit 3D IC performance. In: 2013 IEEE 22nd Conference on Electrical Performance of Electronic Packaging and Systems; 2013. pp. 195-198

[46] Yang C-C, Shieh J-M, Hsieh T-Y, Huang W-H, Wang H-H, Shen C-H, et al. Footprint-efficient and power-saving monolithic IoT 3D+ IC constructed by BEOL-compatible sub-10 nm high aspect ratio (AR>7) single-grained Si FinFETs with record high Ion of 0.38 mA/μm and steep-swing of 65 mV/dec. and I_{on}/I_{off} ratio of 8, in Electron Devices Meeting (IEDM). 2016. pp. 9.1. 1-9.1. 4

[47] Shulaker MM, Wu TF, Pal A, Zhao L, Nishi Y, Saraswat K, et al. Monolithic 3D integration of logic and memory: Carbon nanotube FETs, resistive RAM, and silicon FETs. In: Electron Devices Meeting (IEDM), 2014 IEEE International; 2014. pp. 27.4.1-27.4.4

[48] Shulaker MM, Hills G, Park RS, Howe RT, Saraswat K, Wong H-SP, et al. Three-dimensional integration of nanotechnologies for computing and data storage on a single chip. Nature. 2017;**547**:74-78