The Connection of Neuromorphic Architecture and Bio-Plausible Neurons and Synapses

Sijun Jiang

The Affiliated International School of Shenzhen University, Shenzhen, China, 518000

g421713549@tzc.edu.cn

Abstract. The purpose of this article is to talk about the fundamental functionalilties of the nervous system and its relation to the neuromorphic architecture. The neuromorphic architecture is built to imitate and perform the function of neuron and synapses in our brain. The change phases undergone by neurons and spike timing dependent plasticity performed by synapses are the main reason for the attraction gained in the field of accelerating artificial intelligence and remarkable power efficiency. A neuron is slightly negatively charged because the majority of the anions are interior to the cell membrane while the cations are exterior. Due to the permeability of the cell membrane, ions do not mix. Ions tend to move from region where its concentration is high to region where the concentration is low, causing the phenomena of resting membrane potential. Synapse is the one of the main components responsible for cognitive actions such as learning and forgetting, since it is essential for the transport and the processing of information. Memristors, the material, are used specifically since it has some distinct characteristics that differ from other component to perform the task of neurons and synapses. Different types of memristors are used for different aspects for the tasks, and this article is going to discuss a few of them.

Keywords: Volatile memristor, nonvolatile memristor, neuron, synapse, artificial intelligence, neural network.

1. Introduction

The conventional architecture, von Neumann architecture, has reached its bottleneck of performance to accelerate artificial intelligence. With its tremendous energy and time consumption due to its function logic of processing and storing information, the Von Neumann will become obsolete. Meta, the company owns by Mark Zuckerberg, states that the company in the future will rely heavily on graphics card and will pruchase 350,000 H100 graphics card which will cost over billions of dollars [1]. These graphics cards, specifically the H100 GPU, will consume for loads of energy. The graphic cards, H100, consume so much energy such that each individual will cost for about 700watt of power and 3.5 million of such GPUs will lead to a 13,091,820,000 kilowatt hour energy consumption annually which is approximately the power consumption of some counties in a year (13,092 GWh) [2]. The energy and financial cost of these Von Neumann architecture reflects the problem that it is not suitable for artificial intelligence. In contrary to the power-hungry conventional architecture, our brain does a perfect job at becoming energy-efficient, which only consume 20watt of power. Then an emulation of how our brain works is theorized and given the name, neuromorphic architecture by the man called Carver Mead in the late 1980s.

Compared to the Von Neumann architecture where the information is stored inside the memory units and processed in the CPUs, a neuromorphic architecture will use artificial synapses and neurons to cope with information. Instead of the binary values that serve as the input information for a conventional architecture, spikes serve as the input for a neuromorphic architecture in which the time, magnitude, and shape of the spike could be modelled as parameter that could represent information. There are several fundamental characteristics of a neuromorphic architecture that contributes to its advantages [3].

We study the collocated processing and memories (integrated information storage and calculation). Unlike the conventional architecture where the information is first processed at the CPU and then sent to the memory unit and then sent back, the neuromorphic architecture does not have a separation between neurons and synapses, they both act like the processing and storing units in many of the

implementations. This characteristic specifically facilitates the energy required and therefore resolving the problem of the huge energy requirement of a conventional architecture due to the access from main memory in which needs a sizeable amount of energy.

Next, massive parallelism with the artificial synapses and neurons could all work concurrently and therefor saving time when dealing with repetitive problems. This is a characteristic that a CPU in Von Neumann architecture does not have, in which that CPU has to solve one problem at a time.

These natures of a neuromorphic architecture can deal with the bottleneck of contemporary architectures and specializing in artificial intelligence applications.

Recent advances in technologies reveal that memristors are the most advocated candidates for neuromorphic architecture. The memristors are renowned for its switching speed between high resistance state and low resistance stage, power consumption, density, device stability, and endurance. Both volatile and nonvolatile memristors could be applicated for neuromorphic architecture to emulate the basic functionality of both neurons and synapses [4-11].

Several underlying features for our nervous system are studied and targeted for neuromorphic architecture. Neurons, one of the main components for the system, participate in the action of firing and emitting signals and information. It undergoes through several phases during the process which extremely alters its potential and emits pulses toward to the next neuron. Synapses, the other main components for the nervous system, own the special characteristic of synaptic plasticity. Synaptic plasticity is the variation of the strength of the action potential evoked and is determined by the time function between presynaptic and postsynaptic spikes. These processes are the targets for emulation and goals for neuromorphic architecture [12-17].

2. Method

2.1. Neuron

Neuron is one of the main components of our nervous system which is used as a reference to the neuromorphic architecture. A neuron is constituted of axons, dendrites, and soma in which nucleus and other organelles lie. One typical trait of a neuron is its resting membrane potential. The potential difference on both side of the cell membrane is the underlying cause for the resting membrane potential. The main cause for the resting membrane potential is due to the motion of sodium and potassium ion. The concentration of ions outside the cell over the concentration inside the cell is defined as the concentration gradient. To create a concentration gradient, energy is required. An ATPase pump uses energy to pump 3Na+ ion out of the cell and T the same time pumping 2K+ion into the cell to build a concentration gradient. A concentration gradient is what causes the ions to move. Owing to the semi-permeability characteristic of the cell membrane being permeable to potassium ions but not to sodium ion, potassium ions will move down its concentration gradient going to the extracellular region and leaving anion behind in the nucleus. This movement of charges will cause the cell to exhibit a negative charge and the extracellular region to exhibit a positive charge. This contributes to the built up of an electrostatic gradient. Through time, there will be a force to resist the movement of anions to leave nucleus, which end up with the resting membrane potential.

During the process of transmitting signals, there are three phases for a neuron, depolarization, repolarization and hyperpolarization. The stage depolarization happens when a threshold voltage of the resting membrane is reached. The voltage gated sodium channel of the postsynaptic dendrite will open to allow the inflow of sodium ions. This flow will then depolarize the cell to a larger extent, and therefore opening more voltage gated sodium channel and increasing the potential of the cell body from a negative value to a positive. The next stage is repolarization. After the voltage gated sodium channel is opened for 1 msec, the voltage gated potassium channel will open and allow the inflow of potassium ion. As a consequence, potassium ions will move down their concentration gradient and the potential to decrease to the original negative value. When the membrane potential is below the threshold voltage, voltage gated sodium channel will be immediately closed and the voltage gated potassium channel will close but after voltage gated sodium channel has closed. The period when

only the voltage gated potassium channel is open and the membrane potential falls below threshold potential is called hyperpolarization. As a result, the potential of the cell body will fall below the resting membrane potential. Then the excessive potassium ion will leave the cell through the semipermeable membrane through the activated sodium-potassium pump.

2.2. Synapses

Another main member of the nervous system is the synapse. A synapse is consisted of three components, the presynaptic axon terminal from the presynaptic neuron, the synaptic cleft between presynaptic and postsynaptic neurons, and the postsynaptic dendrite of the postsynaptic neuron. The synapse has a defining characteristic of synaptic plasticity. A synaptic plasticity is defined as the change in the synaptic strength. The spike-timing dependent plasticity is when the time between the presynaptic and postsynaptic spikes determines the forms of plasticity. The plasticity could be categorized in to two forms, Long Term Potentiation (LTP) and Long Term Depression (LTD). During an excitatory synaptic transmission, glutamate is used and acted with the postsynaptic ionotropic glutamate receptor. Ionotropic glutamate receptor can be categorized in to α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA) receptors (AMPAR), N-methyl-D-aspartate(NMDA) receptors (NMDR), and kainate receptors. The role of the learning and memory of our brains is mediated by the AMPARs trafficking in postsynaptic neuron. The AMPARs are permeable to sodium and potassium ion movement and the NMDARs are blocked magnesium and negative voltages. Once the cell is depolarized, the NMDARs are no longer blocked and therefore allowing potassium and calcium ions to move through.

If the presynaptic neuron fires first, it becomes depolarized and release glutamate which will connect to the AMPARs at the postsynaptic neuron. Then the postsynaptic neuron will be depolarized enabling the NMDARs to be permeable and therefore allowing large number of calcium ions reflux.

If the postsynaptic neuron fires first, the postsynaptic neuron will be first depolarized. As the postsynaptic neuron is repolarizing, the presynaptic neuron will fire and release glutamate. The glutamate has less receptor to bind to since the postsynaptic neuron is repolarizing and NMDARs are closed, which leads to a puny amount of calcium ions influx.

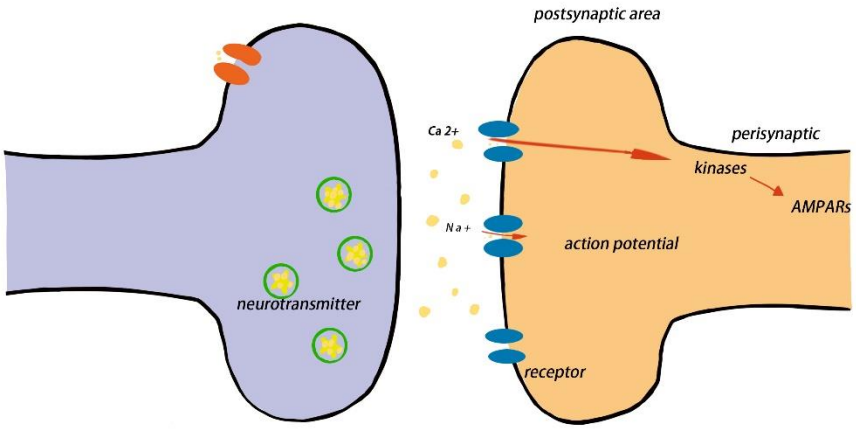


Figure 1. LTP and LDP

In Figure 1, AMPARs interior to the cell are constantly undergoing exocytosis which the AMPARs go from the perisynaptic site to the postsynaptic area. At the postsynaptic area, it will undergo endocytosis to return back to the perisynaptic site and are brought back into the cell in which endosomes carry a large sum of AMPARs. A calcium influx exceeding the threshold value could activate calcium dependent kinases which could alter the process undergone by AMPARs by increasing the amount of exocytosis. This contributes to an increased amount of receptors to bind with when glutamates are released from the presynaptic neuron. Therefore the change in potential will increase and finally contributing to an enhanced connection between neurons and leading to LTP. However, if not enough calcium ions are present, protein phosphatases instead of kinases are released. It also alters the process of AMPARs, but increases the amount of endocytosis in which AMPARs travel from postsynaptic area to perisynaptic areas. Decreased number of receptors at the postsynaptic area means that there are less receptors for glutamate to bind to and therefore leading to weakened synaptic strength or LTD. This intrinsic property has provided a chance for spike-timing-dependent plasticity learning rule algorithm for the acceleration for artificial intelligence.

3. Result

3.1. Memristors

Memristors have excellent features such as power consumption, switching speed, endurance, retention(time taken for the resistance to go back to its original value), and ON/OFF ratios so that it has become one of the most promising candidates for neuromorphic architecture. Memristors in general could be separated into two categories, volatile and nonvolatile. The categories are determined by the memristor's retention time of the low resistance state. Nonvolatile memristors tend to preserve the high resistance state and low resistance, and volatile memristors tend to switch back from low resistance stage to high resistance stage when the voltage applied is removed. These characteristics are determined by the responses, such as thermal effect, ion migration, and magnetic effect, given by the material itself when an external voltage is applied. Different memristors have different switching mechanics. Different input voltage amplitude will also cause different retention time. The features of the switching of states for the volatile and nonvolatile memristors have enabled it to become components for neuromorphic architecture to mimic the behavior of our nervous system, and some memristors even share both characteristic of nonvolatile and volatile memristors.

3.2. Volatile memristors

There are two main switching categories for volatile memristors, digital switching and analog switching. For digital switching category, the current or resistance will boost when the external voltage reaches a threshold voltage, and will decrease sharply when threshold voltage is not reached with a high ON/OFF ratio. For analog switching category, the resistance will change continuously with a low ON/OFF ratio. These threshold voltage switching types of memristors could be very competent at performing works of neurons.

3.3. Resistance-based nonvolatile memristor

The memristor is conventionally two-terminal structured meaning that there will be a top electrode and a bottom electrode with resistive layer in between. The working mechanism for resistance-based nonvolatile memristor tend to focus more on redox reaction. There are several types for the resistance-based nonvolatile memristor, such as valence change memory, electrochemical metallization cell, and phase change memory. For electrochemical metallization cells, the structure will be a tin solid electrolyte switching film stuck in between an active electrode and a counteractive electrode. The active and counteractive electrodes have the characteristic that the metal for active electrode should be able to move in the solid electrolyte film and its redox reaction could happen at a relatively low potential, while the counteractive electrode should be made from metals that are incapable of moving in the film and redox reaction will not be able to perform at the same potential.

The motion and cluster of ions in the thin electrolyte switching film due to the rate of redox reaction and mobility of cations will trigger the formation and deformation of filaments which determine the resistance state. These electrochemical metallization cells are found capable to emulate the dendrite of a neuron.

For the conductive filament memristors, redox reaction will happen triggering the movement of anions and cations leading to the formation and deformation of this conductive filament, and therefore giving the memristor the ability to change between high resistance state and low resistance state. Redox-based memristive devices(ReRAMs) devices belonging to the category conductive filament memristors are found capable of emulating the function performed by neurons. Inside the neuron where the cell membrane separates the ions, a thin solid electrolyte switching film separates two different electrodes with different chemical potentials inside the ReRAMs. The resting membrane potential of a neuron could be emulated by the different electrodes. The role of dendrites and axons is played by the top and bottom electrode and cell is played by the thin film, since the solid electrolyte switching film could sharply alter its conductance due to a small input voltage. When there is an input voltage applied through the top electrode, the potential of the memristor will be built up until a threshold is reached.

Once the threshold is reached, a current will be exerted by the top electrode destinating to the bottom electrode. This specific process precisely emulated how does a neuron fire signal. ReRAMs could be categorized into two groups, non-volatile memristors, specializing in bio-plausible synapses, and volatile memristors.

There are many models designed to mimic the functionality of a neuron such as the Hudgkin-Huxley model, specializing at the field of neuroscience, and the fast integrate and fire model, specializing at the field of artificial intelligence. These bio-plausible artificial neurons and synapses could be applicated into artificial neural networks (ANN). ANNs are separated into spiking neural networks (SNN), and deep neural networks (DNN). A spiking neural network is implemented based on the functionality of spike-time-dependent plasticity in a synapse, in which the resistance of the memristor would change based on the timing function between input and output voltage serving as the weighting function between layers in neural networks. This could be implemented through a crossbar array. These crossbar arrays specifically and specializes in the assembling of bio-plausible neurons and synapses.

4. Conclusion

The characteristic of the functionality of our information processing is desired at the field of accelerating of artificial intelligence and reducing the energy and time cost of the process of information processing. The characteristic of our brain owning this spectacular power and time efficiency is determined by our neurons and synapses. The signal firing process of neurons are emulated. The separation of different cations, sodium and potassium, inside of the soma of neuron has led to the built up of the resting membrane potential due to the negatively charged intracellular body and positively charged extracellular body. The resting membrane potential will react to a threshold voltage built by depolarizing. The depolarization process will lead to the increased potential inside the cell due to inflow of ions. Then the action potential generated by the inflow of ions will travel down the cell body to the next neuron. The repolarization and hyperpolarization will happen to restore the resting membrane potential. The stages of neurons undergoing stages of depolarization, repolarization, and hyperpolarization due can symbolize the information processing and storge. Different types of memristors are used to emulate and concentrate on different aspects of a neuron and its process of firing. Different models are used so that different mechanics, Memristors and goals are oriented. Synapses also possess a property that enable the possibility to accelerate computation of artificial intelligence. The property is such that the strength of connection between neurons can change in the long run due to the time function between the presynaptic and postsynaptic spike. A presynaptic spike followed by a postsynaptic spike could strengthen the connectivity and

given the name LTP. A postsynaptic spike followed by a presynaptic spike could weaken the connectivity between the neurons and given the name LTD. By changing the resistance or the synaptic strength, the connectivity is changed leading to the formation of a weighting function that is controlled and modulated. These characteristics have enabled spike timing dependent plasticity learning rule, as the connectivity can serve as the weighting function between two layers of the neural network.

References

[1] Vanian, J. "Mark Zuckerberg Indicates Meta Is Spending Billions of Dollars on Nvidia AI Chips." *CNBC* (2024).

[2] Shilov, A. "Nvidia's H100 Gpus Will Consume More Power than Some Countries - Each GPU Consumes 700W of Power, 3.5 million Are Expected to Be Sold in the Coming Year." *Tom's Hardware*, Tom's Hardware (2023).

[3] Schuman, C.D., Kulkarni, S.R., et al. "Opportunities for Neuromorphic Computing Algorithms and Applications." *Nature Computational Science*, vol. 2, no. 1, 52 (2021).

[4] Jeong, D.S., and Cheol S.H. "Nonvolatile memory materials for neuromorphic intelligent machines." *Advanced Materials*, vol. 30, no. 42 (2018).

[5] Zhou G., Wang Z., et al. "Volatile and Nonvolatile Memristive Devices for Neuromorphic Computing." *Advanced Materials*, vol 3, 53 (2022).

[6] Choi S., Yang J., and Wang G., "Emerging memristive artificial synapses and neurons for energy-efficient neuromorphic computing." *Advanced Materials*, vol. 32, no. 51, 24 (2020).

[7] Dittmann, R. and Strachan J. P., "Redox-based memristive devices for new computing paradigm." *APL Materials*, vol. 7, no. 11, 244 (2019).

[8] Wang Y., Gong T., et al. "Redox memristors with volatile threshold switching behavior for neuromorphic computing." *Journal of Electronic Science and Technology*, vol. 20, no. 4, 561 (2022).

[9] Cheng W., Song L., et al. "Essential characteristics of memristors for neuromorphic computing." *Advanced Electronic Materials*, vol. 9, no. 2, 3 (2022).

[10] Wang R., Yang J., et al. "Recent advances of volatile memristors: Devices, mechanisms, and applications." *Advanced Intelligent Systems*, vol. 2, no. 9, 22 (2020).

[11] Chen S., Zheng T., et al. "Electrochemical-memristor-based artificial neurons and synapses—fundamentals, applications, and challenges." *Advanced Materials*, vol. 35, no. 37, 569 (2023).

[12] Rider M. H., Jessu R, K. R. "Physiology, Action Potential." Treasure Island (FL): StatPearls Publishing (2024).

[13] Chrysafides SM, Bordes SJ, Sharma S. "Physiology, Resting Potential." Treasure Island (FL): StatPearls Publishing (2024).

[14] Lee HK. "AMPA Receptor Phosphorylation in Synaptic Plasticity: Insights from Knockin Mice." In: Kittler JT, Moss SJ, editors. The Dynamic Synapse: Molecular Methods in Ionotropic Receptor Biology. Boca Raton (FL): CRC Press/Taylor & Francis (2006).

[15] Sumi, T. and Kouji H. "Mechanism underlying hippocampal long-term potentiation and depression based on competition between endocytosis and exocytosis of AMPA receptors." *Scientific Reports*, vol. 10, no. 1, 34 (2020).

[16] Díaz A. J., and Roger A. N. "AMPA receptor trafficking and LTP: Carboxy-termini, amino-termini and tarps." *Neuropharmacology*, vol. 197, 55 (2021).

[17] Magee, Jeffrey C., and Christine G. "Synaptic plasticity forms and functions." *Annual Review of Neuroscience*, vol. 43, 1 (2020).