Chapter

Opening the "Black Box" of Silicon Chip Design in Neuromorphic Computing

Kangjun Bai and Yang Yi

Abstract

Neuromorphic computing, a bio-inspired computing architecture that transfers neuroscience to silicon chip, has potential to achieve the same level of computation and energy efficiency as mammalian brains. Meanwhile, three-dimensional (3D) integrated circuit (IC) design with non-volatile memory crossbar array uniquely unveils its intrinsic vector-matrix computation with parallel computing capability in neuromorphic computing designs. In this chapter, the state-of-the-art research trend on electronic circuit designs of neuromorphic computing will be introduced. Furthermore, a practical bio-inspired spiking neural network with delay-feedback topology will be discussed. In the endeavor to imitate how human beings process information, our fabricated spiking neural network chip has capability to process analog signal directly, resulting in high energy efficiency with small hardware implementation cost. Mimicking the neurological structure of mammalian brains, the potential of 3D-IC implementation technique with memristive synapses is investigated. Finally, applications on the chaotic time series prediction and the video frame recognition will be demonstrated.

Keywords: analog signal processors, lab on a chip, neuromorphic computing, reservoir computing, analog/mixed-signal circuit design, three-dimensional integrated circuit, image classification

1. Introduction

Benefit by the Moor's law, the von Neumann computing architecture, respectively storing and processing data instructions in the memory unit and the central processing unit (CPU), was served as the major computing model in past several decades [1]. However, physical limitations of the complementary metal-oxide-semiconductor (CMOS) technology and the storage capacity hinder the performance development of classic computers; such classic computers can no longer double its performance every 18 months, indicating the end of Moore's prediction [2].

Recently, the computing efficiency of extracting valuable information in data-intensive applications through the von Neumann computing architecture has become computationally expensive, even with super-computers [3]. The accumulated amount of energy required for the data processing through super-computers poses a query on whether the augmented performance is sustainable.

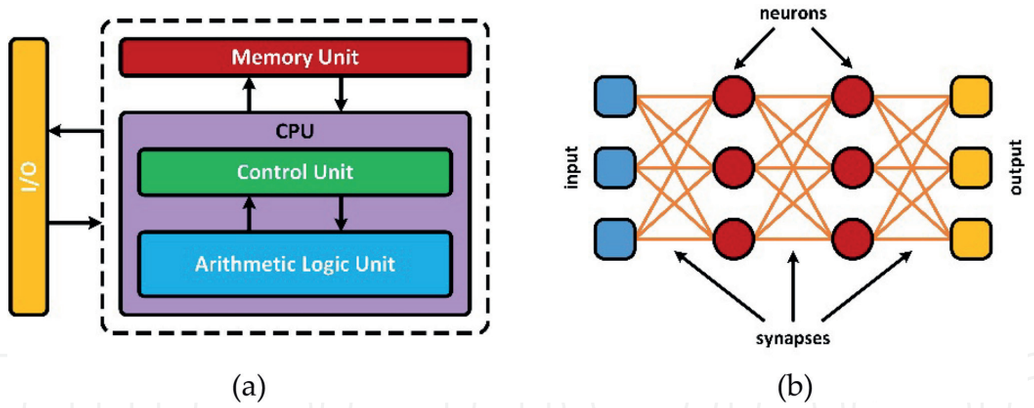

Figure 1.
General architecture of (a) von Neumann computing system and (b) neuromorphic computing system.

As human beings, our brains are capable to analyze and memorize sophisticated information with only $20W$ of energy consumption [4]. In the 1980s, neuromorphic computing, proposed by Dr. Carver Mead, has matured to provide intelligent systems that able to mimic biological processes of mammalian brains through highly parallelized computing architectures; such systems typically model the function of neural network through very-large-scaled-integrated (VLSI) circuits [5]. Major differences between the von Neumann computing architecture and the neuromorphic computing system are illustrated in **Figure 1**. Recently, artificial neural networks (ANNs) have demonstrated their superior performance in many data-extensive applications, including image classification [6–8], handwritten digit recognition [9–11], speech recognition [12, 13] and others. For instance, *TrueNorth*, the neuromorphic chip fabricated by IBM in 2014, is capable to classify multiple objects within a 240 × 240-pixel video input with merely $65mW$ of energy consumption. Compared to the von Neumann computing system, such a neuromorphic computing system has five orders of magnitude more energy efficient [14]. *Loihi*, the latest prototype of brain-inspired chip fabricated by Intel in 2017, involves a mere 1/1000 power consumption of the one used by a classic computer [15].

Most recent hardware implementations on neuromorphic computing systems focus on the digital computation because of its advantages in noise immunity [16]. However, real-time data information is often recorded in the analog format; thereby, power-hungry operations, such as analog-to-digital (A/D) and digital-to-analog (D/A) conversions, are needed to facilitate the digital computation. It can be observed that the digital computation results in high power consumption with a large design area.

In this chapter, an overview of ANNs will be discussed in Section 2. Section 3 introduces the spiking information processing technique through the temporal code with the leaky integrate-and-fire neuron. Our fabricated spiking neural network chip along with its measurement results on the chaotic behavior will be demonstrated in Section 4, followed by the investigation on 3D-IC implementation technique with memristive synapses in Section 5. Applications on the chaotic time series predication and the image recognition are illustrated in Section 6.

2. Artificial neural networks

In the endeavor to imitate the nervous system within mammalian brains, ANNs are built by employing electronic circuits to imitate biological neural networks [17]. In general, ANN methodologies adopt the biological behavior of neurons and synapses, so-call the hidden layer, in their architecture. The hidden layer is constituted by multiple "neurons" and "synapses", which carries activation functions that

control the propagation of neuron signals. Based on the connection pattern and the learning algorithm, ANN methodologies can be classified into various categories, as depicted in **Figure 2**.

The multilayer perceptron (MLP), a representation of feedforward neural networks (FNNs), is composed by unidirectional connections between hidden layers. MLP has become the quintessential ANN model due to its advantages in ease of implementation [18]. However, the major design challenge in the MLP is that the runtime as well as the training and learning accuracy of the system are strongly affected by the number of neurons and hidden layers. As the neural information evolved into a much more sophisticated mixed-signal evaluation, disadvantages of MLP are exposed when such a neural network is deployed for temporal-spatial information processing tasks [19]. Recurrent neural networks (RNNs), successfully adopt the temporal-spatial characteristics within their hidden layer, closely mimic the working mechanism of biological neurons and synapses. However, the major design challenge is that all weights within the network need to be trained, which dramatically increases its computational complexity. In earlier 2000s, the reservoir computing, an emerging computing paradigm, exploits the dynamic behavior of conventional RNNs and computationally evolved its training mechanism [20]. Within the reservoir layer, synaptic connections are constructed by a layer of nonlinear neurons with fixed and untrained weights. In the reservoir computing, the complexity of the training process is significantly reduced, since only output weights are needed to be trained, thereby, higher computational efficiency can be achieved.

The conventional reservoir computing has been fully developed in the past decade to simplify the training operation of RNNs and proven its benefits across multifaceted applications [21–24]; however, the computational accuracy of the system is still highly proportional to the number of neurons within the reservoir layer. It can be observed that these enormous numbers of neurons significantly hinder the hardware development on the reservoir computing. In [25], it has been proven that the computing architecture is capable to exhibit rich dynamic behaviors during operations when the delay is employed into the system. Benefit from the embedded delay property, the training mechanism and the computing architecture of conventional reservoir computing have conceptually evolved, namely the time delay reservoir (TDR) computing [26]. In the TDR computing, the reservoir layer is built by only one nonlinear neuron with a feedback loop. In this context, time-series input data can be processed through the TDR computing by taking advantages of the feedback signal to form a short-term memory, thereby, higher computational efficiency and accuracy can be achieved.

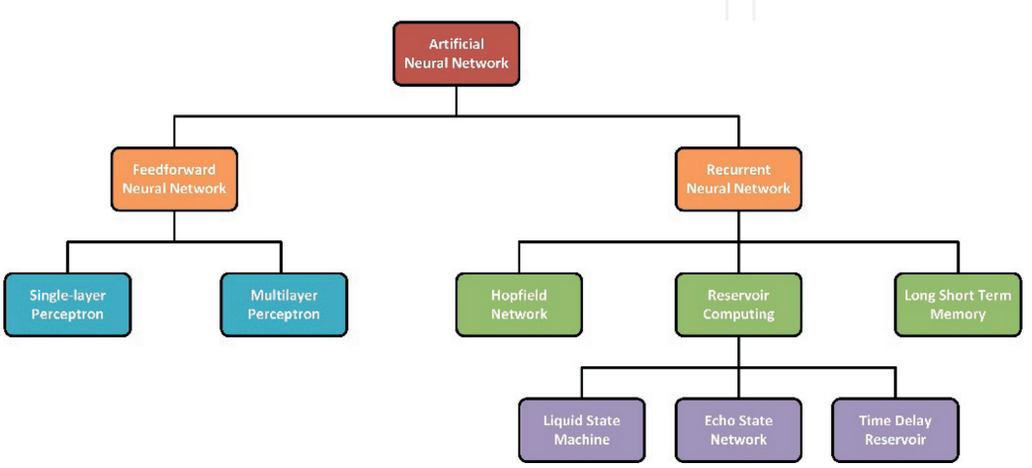

Figure 2.
Overview of artificial neural networks.

3. Spiking information processing

In many brain-inspired neuromorphic computing systems, the interface between modules is often influenced by the signal propagation. The major design challenge in neuromorphic computing is the difficulty in adapting raw analog signals into a suitable data pattern, which can be used in the neuronal activities. Before digging deep into the architecture of our fabricated spiking neural network chip, in this section, a temporal encoding scheme through the analog IC design technique will be discussed.

3.1 CMOS neuron models

In past few decades, researches on biological neurons have been fully investigated in the field of neuroscience [27–32]. In general, the dendrite, the soma, the axon and the synapse are four major elements of a biological neuron [33]. Within a nervous system, dendrites collect and transmit neural signal to the soma, while the soma plays an important role as the CPU to carry out the operation of the nonlinear transformation. Moreover, signals are processed and transmitted in form of a nerve impulse, also known as the spike [34]. During the operation, an output spike is formed when the input stimulus surpasses the threshold level, indicating as the firing process. **Figure 3** demonstrates a typical firing and resting operation in a biological neuron. Synapses along with the axon are then transmitted the spike data patterns to other neurons.

The leaky integrate-and-fire (LIF) neuron model plays an important role in the neuron design to convert raw analog signals into spikes [35]. **Figure 4** depicts the analog electronic circuit model of a LIF neuron. The input excitation, I_{ex}, can be expressed as

$$I_{ex} = C_m \cdot \frac{d V_m}{dt} + I_{leak}, \tag{1}$$

where C_m is the membrane capacitance, $\frac{d V_m}{dt}$ represents the voltage potential across the membrane capacitor over time, and I_{leak} is the leakage current. During the operation, raw analog signals are firstly converted into an excitation current, which will be used to charge up the potential level across the membrane capacitor. When the voltage potential across the membrane capacitor surpasses the threshold level, the circuit fires a spike as its output. Once the firing process is accomplished, the membrane capacitor will be reset to its initial state until the next firing cycle takes

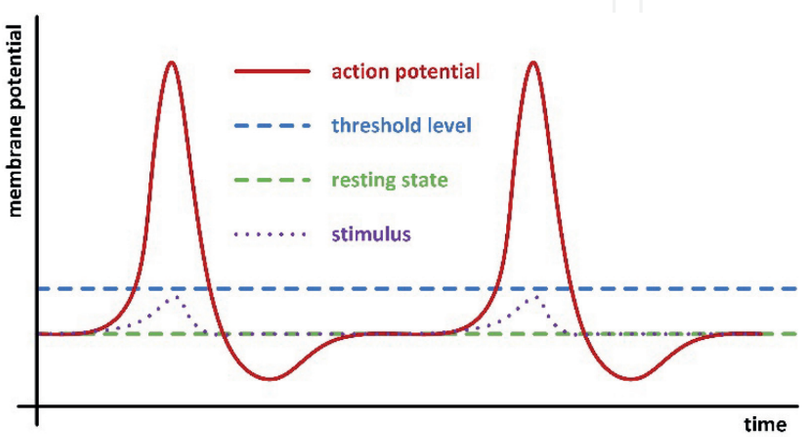

Figure 3.
Action potential of biological impulses.

DOI: http://dx.doi.org/10.5772/intechopen.83832

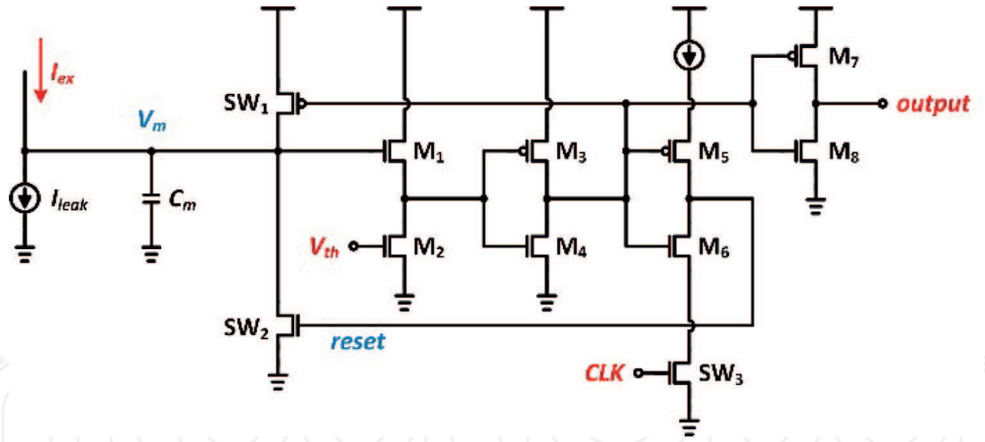

Figure 4.
Analog electronic circuit model of a LIF neuron.

place. The LIF neuron is capable to process both firing and resetting operations, closely mimicking the biological behavior of neurons.

From Eq. (1), it can be observed that the integration time over the membrane capacitor can be regulated by excitation and leakage currents. Such relation can be depicted by a simple resistor model, which can be rewritten as

$$I_{ex} = C_m \cdot \frac{d\,V_m}{dt} + \frac{V_m}{R_{leak}}, \tag{2}$$

where $\frac{V_m}{R_{leak}}$ determines the amount of leakage current. Thereby, the voltage potential across the membrane capacitor can be determined as

$$V_m = I_{ex} \cdot R_{leak} - e^{\frac{t}{R_{leak} \cdot C_m}}. \tag{3}$$

3.2 Neural codes

Neural code is used to characterize raw analog signals into neural responses. In general, there are two distinct classes to represent neural codes. One class converts analog signals into a spike train where only the number of spikes matters, knowing as the rate code. Another class converts analog signals into the temporal response structure [36] where time intervals matters, knowing as the temporal code.

Figure 5 demonstrates major differences between the rate code and the temporal code. In the rate code, analog signals are encoded into the firing rate within a sampling period, as shown in **Figure 5a**. Considering the implementation complexity, the rate encoding scheme is easier to implement through electronic circuits compared to the temporal encoding scheme; however, small variation of an analog signal in the temporal response structure are neglected, which makes the rate

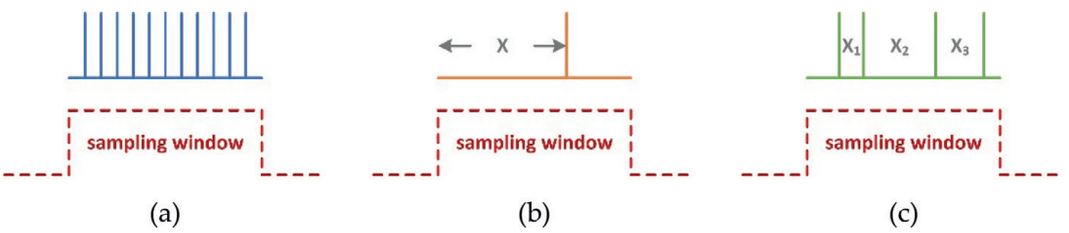

Figure 5.
Neural codes in (a) rate code, (b) time-to-first-spike latency code, and (c) inter-spike-interval temporal code.

code inherency ambiguous in the real-time computation [36]. In [37], researches discover that neural information does not only depend on the spatial, but also the temporal structure. Time-to-first-spike (TTFS) latency code [38–40] is one of the simplest temporal encoding schemes. As demonstrated in **Figure 5b**, in a TTFS latency code, analog signals are encoded into a time interval between the starting point of the sampling period and the generated spike. However, the encoding error would be large if the system performs abnormally.

The inter-spike-interval (ISI) code is another branch of the temporal code, where encoded analog signals depends on the internal time correlation between spikes [41, 42], as illustrated in **Figure 5c**. In general, the ISI temporal encoder converts all analog signals into several inter-spike-intervals, allowing each spike to be the reference frame to others. Obviously, the ISI code is capable of carrying more information within a sampling period compared to the TTFS latency code.

Figure 6a demonstrates the simplified function diagram of ISI temporal encoder. The ISI temporal encoder employs an iteration architecture such that each LIF neuron operates in separate clock periods. The signal regulation layer is built by a current mirror array to duplicate the input excitation current for each LIF neuron; the neuron pool along with the signal integration layer achieve the iterative characteristic. Our ISI temporal encoder chip was fabricated through the standard GlobalFoundries (GF) 180 nm CMOS technology, as depicted in **Figure 6b**.

The number of spikes in an ISI code as discussed in [32] is directly proportional to the number of neurons. Even though this linear proportional correlation is desirable, its hardware implementation is still far more challenging. On the other hand, it can be observed that the exponential relation would increase the number of spikes, thus, containing more information even with the same number of neurons. Through the iterative structured ISI temporal encoder, the number of generated spikes, S_N, can be determined by the number of neurons, which can be written as

$$S_N = 2^N - 1, \tag{4}$$

where N defines the total number of neurons.

From Eq. (4), it can be observed that even with the same number of neurons, the ISI temporal encoder is capable to produce more spikes compared to [35]; thereby,

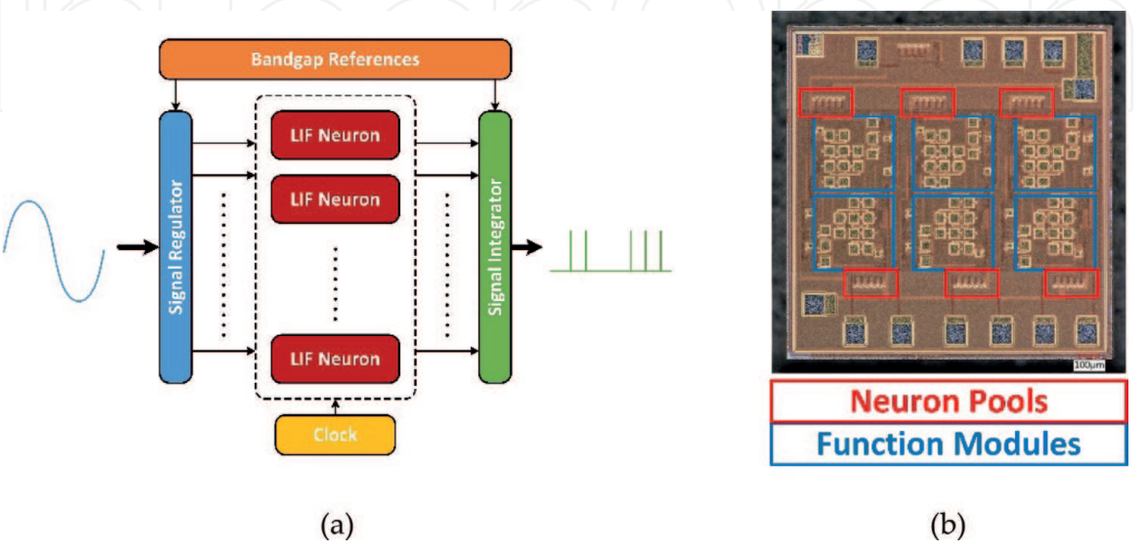

Figure 6.
(a) Simplified function diagram of ISI temporal encoder and (b) die photo of our fabricated ISI temporal encoder chip [32].

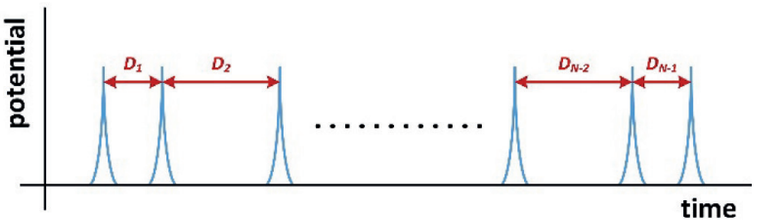

Figure 7.
ISI temporal spike train with N LIF neurons.

the ISI temporal encoder has capability to carry more information. The iterative structure greatly reduces the power consumption, since a smaller number of neurons are needed to produce the equal number of spikes.

In this iterative structured design, the ISI temporal encoder samples the original analog signal without using A/D and D/A conversions, and converts analog signals into several inter-spike-intervals. The expression of the inter-spike-interval can be simplified as

$$D_i = \frac{A_i}{I_{ex} - I_{leak}}, \tag{5}$$

where $A_i = C_m \cdot V_m$. In the IC implementation, the membrane capacitor is fixed, thus, V_i is a constant; thereby, the variable, A_i, in terms of excitation current can be defined as

$$A_i = \beta \cdot A_{N-1} = \beta^2 \cdot A_{N-2} = \cdots = \beta^{N-1} \cdot A_1, \tag{6}$$

where β is an arbitrary design parameter.

The general expression of each inter-spike-interval, as demonstrated in **Figure 7**, can be written as

$$D_{2^{N-1}-1} = \frac{1}{A_N} \cdot \frac{V_{N-1}}{\beta^{N-1}}, \tag{7}$$

$$D_{2^{N-1}-2} = \frac{1}{A_1} \cdot \left(\frac{V_{N-2}}{\beta^{N-2}} - \frac{V_{N-1}}{\beta^{N-1}} \right), \tag{8}$$

$$\vdots$$

$$D_{2^{N-1}} = \frac{1}{A_1} \cdot \left(\frac{V_1}{\beta^1} - \frac{V_2}{\beta^2} - \frac{V_3}{\beta^3} - \cdots - \frac{V_{N-3}}{\beta^{N-3}} - \frac{V_{N-2}}{\beta^{N-2}} - \frac{V_{N-1}}{\beta^{N-1}} \right). \tag{9}$$

4. CMOS nervous system design

With the respect to the analog design of neural code, our spiking neural network chip adapts the ISI temporal encoding scheme as it pre-signal processing module, as well as the reservoir computing module with delay topology as the processing element. Our spiking neural network, named as the analog delayed feedback reservoir (DFR) system is considered as the simplification of conventional reservoir

computing. By employing the delayed feedback structure within the system, our analog DFR system processes the functionality of high dimensional projection and short-term dynamic memory, whereby the behavior of biological neuron is achieved.

4.1 Architecture of analog DFR system

Figure 8 demonstrates the architecture of our analog DFR system, as published in [43, 44]. During the operation, the high dimensional projection within the reservoir layer, as illustrated in **Figure 9**, is the key module to separate input patterns into different categories [26]. For instance, with low dimensional spaces, two different objects cannot be linearly separated by a single cut-off line, as shown in **Figure 9a**. However, by projecting input patterns onto higher dimensional spaces, from two-dimensional to three-dimensional, the separability of the system changes accordingly. As demonstrated in **Figure 9b**, the same objects are linearly separated by a single cut-off plane without changing their original xy position. Our analog DFR chip was fabricated through the GF 130nm CMOS technology, as demonstrated in **Figure 10**.

In our analog DFR system, the dynamic behavior can be controlled by changing the total delay time within the feedback loop. Along the feedback loop, the total delay time, T, is separated into N intermediate neurons with an identical delayed time constant, τ_{delay}, such that

$$\tau_{delay} = \frac{T}{N}. \tag{10}$$

In the conventional reservoir computing system, represented by the echo state network (ESN), the memory within the reservoir layer fades in time due to the way that neurons are sparsely connected; such fading memory limits the performance of computation [20]. With the delay-feedback topology embedded, our analog DFR system not only reduces the implementation complexity but also overcomes the drawback of fading memory limitation. Such functionality enables the knowledge transfer processing technique, allowing new incoming input data to carry information from its previous states, as depicted in **Figure 11**. The expression of N^{th} output, S_N, can be simplified as

$$S_N = f\left[I_p(x) + \sum_{x=1}^{N} I_{P-1}(x) \cdot A v^x\right], \tag{11}$$

where the function, $f_{[\,]}$, represent the nonlinear transformation of input signal; $I_p(x)$ and $I_{p-1}(x)$ indicate the current and previous input patterns, respectively; Av is the finite gain of the gain regulator within the reservoir layer.

4.2 Delay characteristic

Along the feedback loop, the delay time constant, τ_{delay}, can be controlled by the integration time over the membrane capacitor, which can be expressed as

$$\tau_{delay} = C_m \cdot \frac{V_m}{I_{ex}}. \tag{12}$$

In general, the mathematical model of the delay time constant is represented by the values of resistance and capacitance. In the LIF delay neuron, the input impedance, R_{in}, is equivalent to $\frac{V_m}{I_{ex}}$, thus, the delay time constant can be simplified as

$$\tau_{delay} = C_m \cdot R_{in}. \tag{13}$$

The feedback loop, which is constructed by multiple LIF neurons, as illustrated in **Figure 12**. To enable the spiking signal propagation, the output spike train from the previous neuron is utilized as the clock signal to trigger its following neuron.

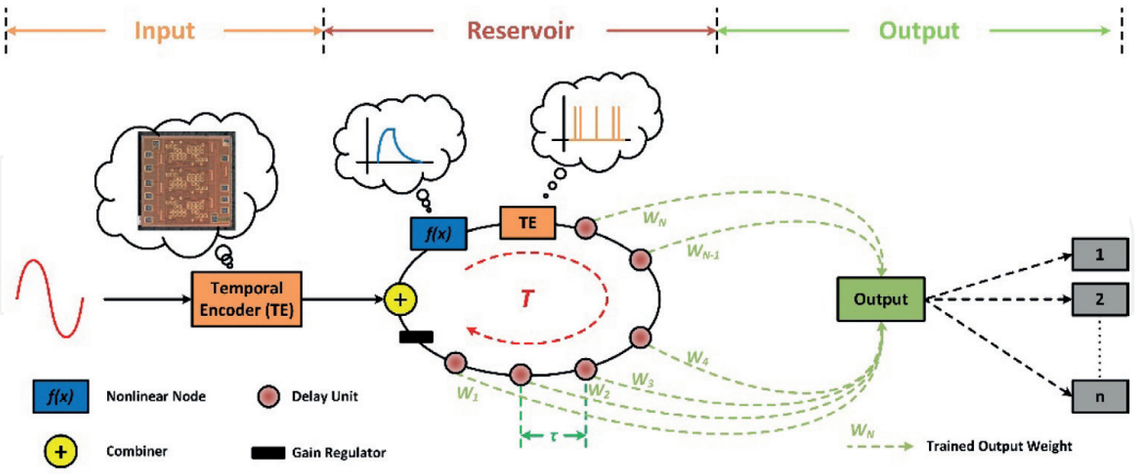

Figure 8.
Architecture of our analog DFR system.

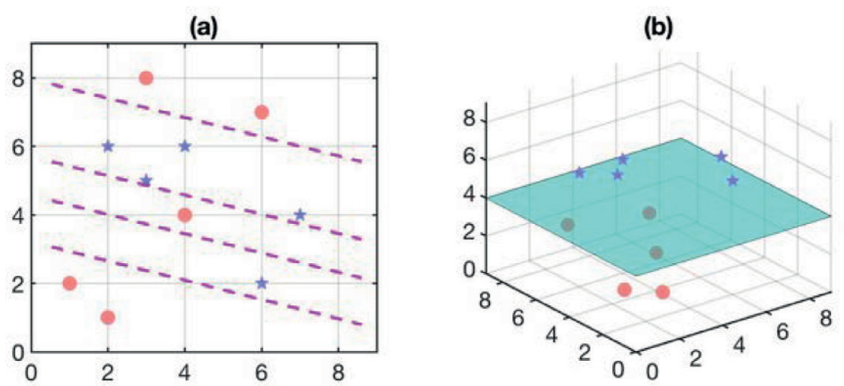

Figure 9.
(a) Nonlinear classification with low dimensional spaces and (b) linear classification with high dimensional spaces.

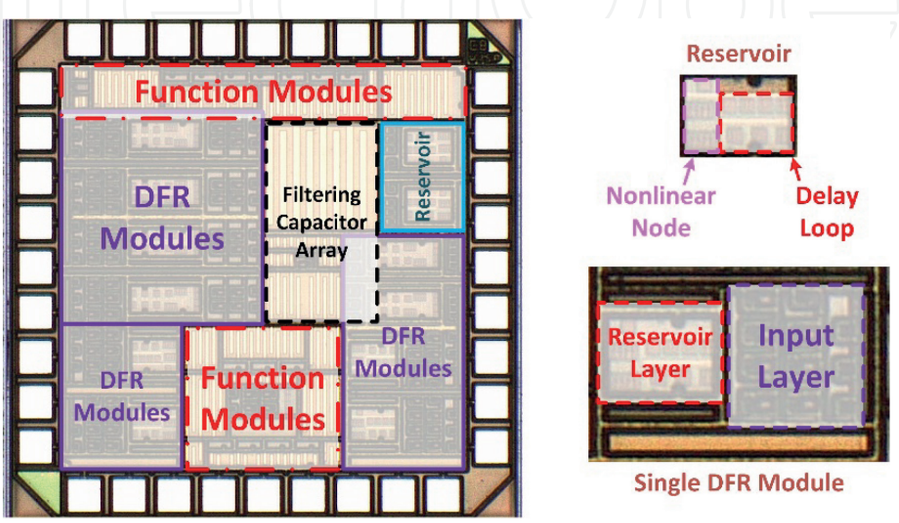

Figure 10.
Die photo of our fabricated analog DFR chip.

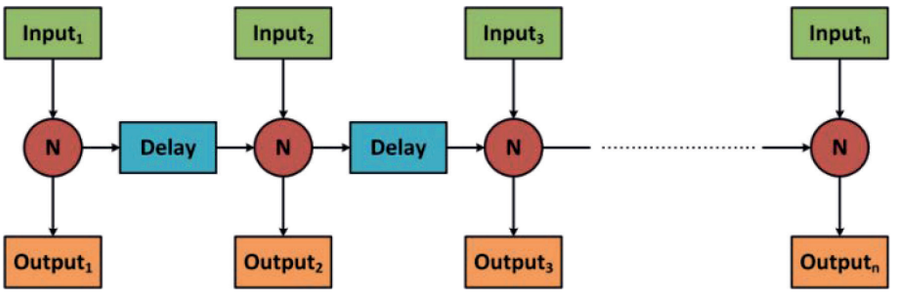

Figure 11.
Illustration of short-term dynamic memory.

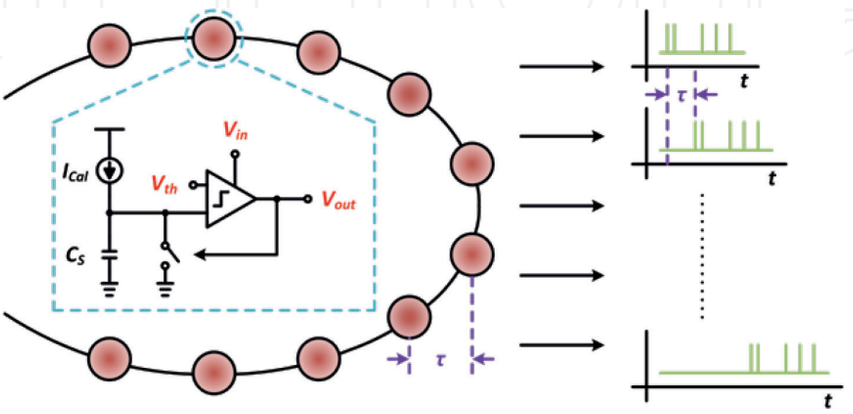

Figure 12.
Dynamic delayed feedback loop.

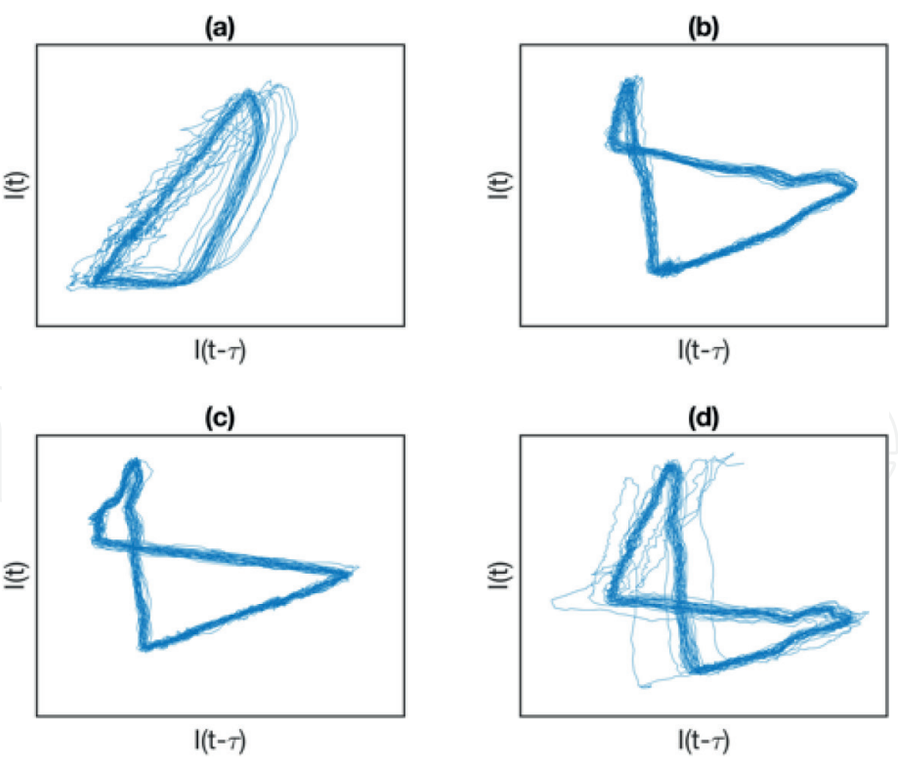

Figure 13.
Measured phase portrait of dynamic system in (a) T = 0.64 μs; (b) T = 1 μs; (c) T = 1.2 μs; (d) T = 1.4 μs.

4.3 Dynamic behavior

In general, the phase portrait is used to visualize how solutions of a delay system would behave. In this experiment, measured phase portraits are plotted through two signals from the feedback loop where one of them is recorded with the time

delay, as shown in **Figure 13**. As the total delay time within the feedback loop increases, the dynamic behavior of the system changes accordingly. As plotted in **Figure 13b**, the delayed signal repeatedly traces its initial path when the total delay time within the feedback loop maintains around 1 μs, indicating as the periodic. When the total delay time within the feedback increases to 1.4 μs, as shown in **Figure 13d**, the delayed signal diverges its initial path but still tracking its equilibrium point, indicating as the edge-of-chaotic.

5. Three-dimensional neuromorphic computing

To closely mimic functionalities of mammalian brains, electronic neurons and synapses in neural network designs need to be constructed in a network configuration, which demands extremely high data communication bandwidth between neurons and high connectivity neural network degree [45, 46]. However, these requirements are not achievable through the traditional von Neumann architecture or the two-dimensional (2D) IC design methodology. Recently, a novel 3D neuromorphic computing system that stacks the neuron and synapse vertically has been proposed as a promising solution with lower power consumption, higher data transferring rate, high network degree, and smaller design area [47, 48]. There are two 3D integration techniques that can be used in the hardware implementation of neuromorphic computing: (1) through-silicon via (TSV) 3D-IC and (2) monolithic 3D-IC. A well-known 3D integration technique is to use the TSV as vertical connection to bond two wafers. In this structure, a large capacitance that is formed by TSVs can be used to build the membrane capacitor, which is required in neuron firing behavior [49–51]. Unlike the TSV 3D-IC technique that uses separately fabrication processes, the monolithic 3D-IC technique is capable to integrate multiple layers of devices at a single wafer, thus, the monolithic 3D-IC technique is capable to provide a smaller design area with lower power consumption [52, 53].

5.1 Memristor

In neural network designs, the electronic circuit model of synapses can be implemented by an emerging non-volatile device, namely the memristor, which is a class of the resistive random-access memory (RRAM). In general, the memristor device is constructed in a metal-insulator-metal (MIM) structure, as illustrated in **Figure 14a**. The resistance of a memristor device can be gradually changed between its low resistance state and high resistance state as the voltage across the memristor device changes.

Memristors are typically fabricated in a 2D crossbar structure [54], which can be further extended to 3D space, as illustrated in **Figure 14c** and **d**, respectively.

5.2 Memristor-based 3D neuromorphic computing

In the field of ANN designs, a novel 3D neural network architecture, which combines memristors and the monolithic 3D-IC technique, has been proposed [55]. In this structure, neurons and memristor-based synaptic array are stacked vertically, as demonstrated in **Figure 15** [48]. As a non-volatile device, RRAM is capable save static power consumption with small implementation area while maintaining its weighted value. With the monolithic 3D-IC technique, the memristor-based 3D neuromorphic computing can potentially reduce the length of critical path by 3X [56], increase the scalability [52], decrease the power consumption by 50% as well as minify the die area by 35% [57].

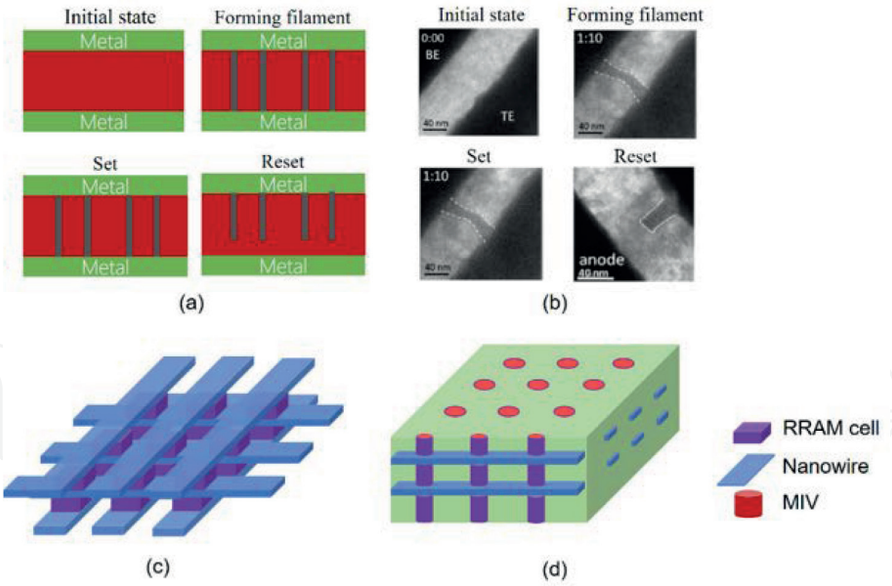

Figure 14.
(a) Switching process of memristor device; (b) transmission electron microscopy images of dynamic evolution of conductive filaments; (c) horizontal RRAM structure; and (d) vertical RRAM structure.

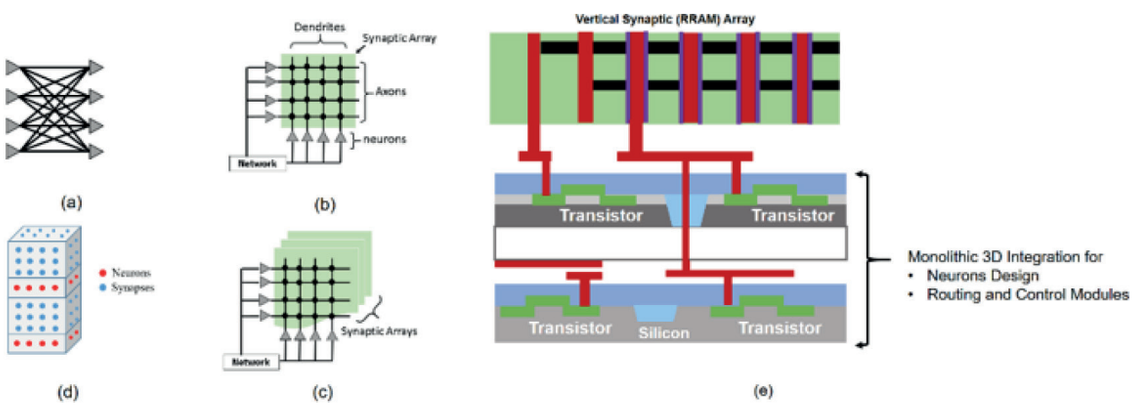

Figure 15.
(a) Typical ANN; (b) 2D structured crossbar array; (c) 3D structured crossbar array; (d) 3D neuromorphic computing architecture by stacking synapses vertically; and (e) deploy monolithic 3D neuromorphic computing on a silicon chip.

6. Lab on a chip

6.1 Chaotic time series prediction

To evaluate the precision of our analog DFR system, a chaotic time series prediction benchmark, the tenth-order nonlinear autoregressive moving average system (NARMA10), is carried out, which can be governed by the following equation

$$O(t) = 0.3 \cdot O(t) + 0.05 \cdot O(t) \cdot \sum_{i=0}^{9} O(t-i) + 1.5 \cdot D(t-9) \cdot D(t) + 0.1, \tag{14}$$

where $D(t)$ is the random input signal at time t, and $O(t)$ is the output signal. In this experiment, 10,000 sampling points were generated through Eq. (14) for training and testing phases. 6000 samples were used for the training while rest samples were used for the testing. The prediction error was then examined through the normalized root mean square error (NRMSE).

In the training phase, output weights were trained by minimizing the deviation between target and predicted outputs. Both training and testing errors were achieved by the NRMSE, which can be defined as

$$NRMSE = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N\sigma_{\hat{y}}^2}}, \tag{15}$$

where y_i defines the predicted output, $\hat{y}_i$ is the target output, N is the number of samples, and $\sigma_{\hat{y}}^2$ determines the output variance. Experimental results of predicted output signals against target outputs with our analog DFR computing system is plotted in **Figure 16**. From experimental results, the training and testing errors are found to be 8.49 and 6.83%, respectively.

6.2 Video frame recognition

In this task, the application of video frame recognition is chosen to examine the performance of our analog DFR system. In this experiment, 48 images, which comprise three different persons with various face angles, were drawn from the Head Pose Image dataset [58], as demonstrated in **Figure 17a**. Twenty images were used for the training, while another 24 images were used for the testing. In the training phase, the face angle changes from 0 to 75° horizontally. In the testing phase, the rotational angle of face follows the training phase but with additional 15° applied vertically.

As illustrated in Section 4.3, our fabricated analog DFR chip is capable to operate at the edge-of-chaos region as the delay changes. To demonstrate the importance of delay, our model was evaluated through several delayed time constants. As depicted in **Figure 18**, it can be observed that the recognition rate changes with regard to the delay time. For instance, the recognition rate maintains above 98% when the system operates at the edge-of-chaos regime (T = 20 ms) with 10% or less salt-and-pepper noise. As the noise level approaches to 50%, the recognition rate still maintains above 93%. However, if the dynamic behavior of the system deviates from the edge-of-chaos regime, the recognition rate significantly reduces due to the change in the dynamic behavior.

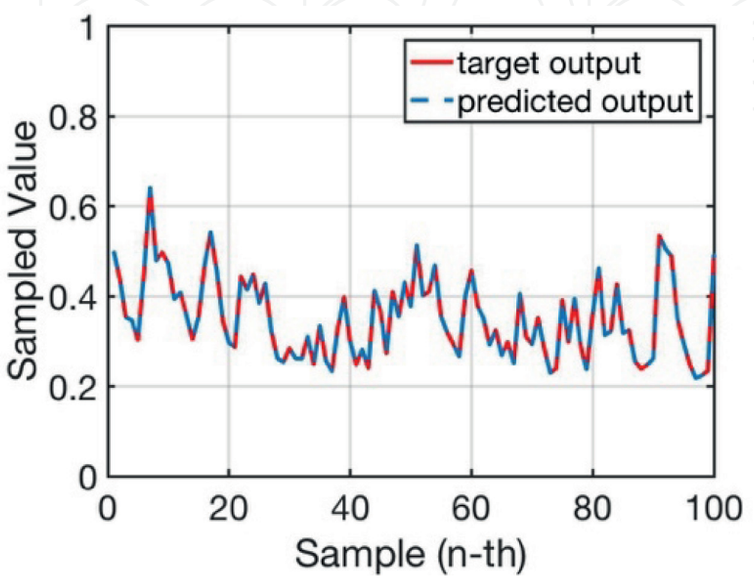

Figure 16.
Target signals versus predicted signals for NARMA10 benchmark.

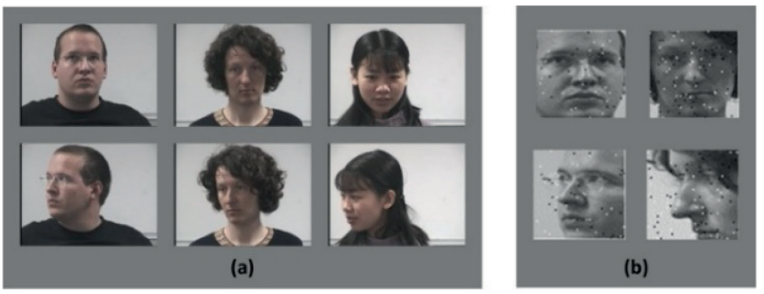

Figure 17.
(a) Training database with three subjects and (b) testing dataset with various salt-and-pepper noise levels.

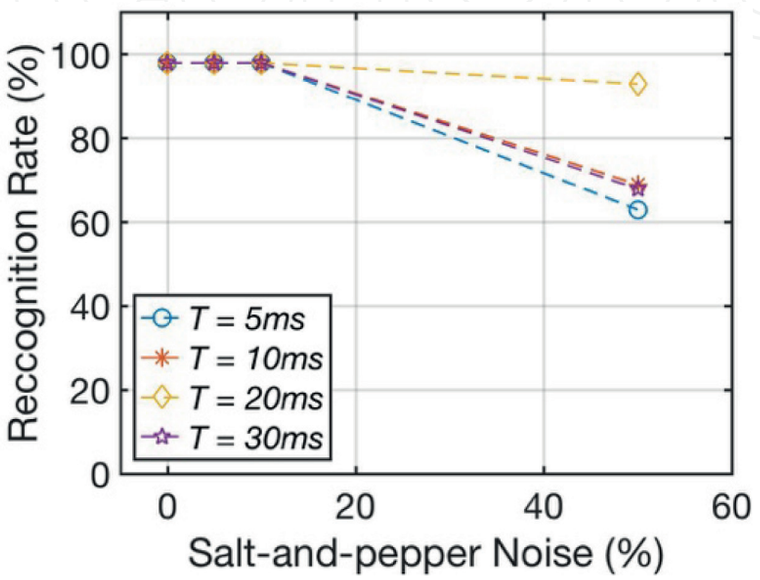

Figure 18.
Recognition rate with respect to various dynamic behavior.

7. Conclusions

In this chapter, the design aspect of our analog DFR system with the analogue electronic circuit model of biological neuron is discussed. By mimicking how human beings process information, our analog DFR system adapts the spiking temporal information processing technique and a nonlinear activation function to project input patterns onto higher dimensional spaces. From measurement results, our analog DFR system demonstrates richness in dynamic behaviors, closely mimicking the biological neurons with delay property. By naturally perform these neuron-like operations, our analog DFR system is capable to nonlinearly project input patterns onto higher dimensional spaces for the classification while operating at the edge-of-chaos region with merely 526 μW of power consumption. Experimental results on the chaotic time series prediction and the video frame recognition demonstrate the high recognition accuracy even with noise, making our analog DFR system a candidate for low power intelligence applications.

DOI: http://dx.doi.org/10.5772/intechopen.83832

IntechOpen

Author details

Kangjun Bai and Yang Yi*
The Bradley Department of Electrical and Computer Engineering, Virginia Polytechnic Institute and State University, Blacksburg, Virginia, USA

*Address all correspondence to: yangyi8@vt.edu

References

[1] Von Neumann J. The Computer and the Brain. New Haven, USA: Yale University Press; 2012

[2] Schaller RR. Moore's law: Past, present and future. IEEE Spectrum. 1997;**34**(6):52-59

[3] Dayarathna M, Wen Y, Fan R. Data center energy consumption modeling: A survey. IEEE Communication Surveys and Tutorials. 2016;**18**(1):732-794

[4] Yu S et al. An electronic synapse device based on metal oxide resistive switching memory for neuromorphic computation. IEEE Transactions on Electron Devices. 2011;**58**(8):2729-2737

[5] Mead C. Neuromorphic electronic systems. Proceedings of the IEEE. 1990;**78**(10):1629-1636

[6] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems (NIPS). 2012. pp. 1097-1105

[7] Jia Y et al. Caffe: Convolutional architecture for fast feature embedding. In: Proceedings of the 22nd ACM international conference on Multimedia. Florida, USA: ACM; 2014

[8] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556; 2014

[9] LeCun Y et al. Handwritten digit recognition with a back-propagation network. In: Advances in Neural Information Processing Systems (NIPS). 1990. pp. 396-404

[10] Bottou L et al. Comparison of classifier methods: A case study in handwritten digit recognition. In: IEEE Proceedings of the 12th IAPR International. Conference on Pattern Recognition, 1994. Vol. 2. Conference B: Computer Vision & Image Processing. 1994

[11] Cireşan DC et al. Deep, big, simple neural nets for handwritten digit recognition. Neural Computation. 2010;**22**(12):3207-3220

[12] Hinton G et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal Processing Magazine. 2012;**29**(6):82-97

[13] Mikolov T et al. Strategies for training large scale neural network language models. In: 2011 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU). Hilton Waikoloa Village, Waikoloa Village, HI: IEEE; 2011

[14] Merolla PA et al. A million spiking-neuron integrated circuit with a scalable communication network and interface. Science. 2014;**345**(6197):668-673

[15] Davies M et al. Loihi: A neuromorphic manycore processor with on-chip learning. IEEE Micro. 2018;**38**(1):82-99

[16] Eryilmaz SB et al. Neuromorphic architectures with electronic synapses. In: 2016 17th International Symposium on Quality Electronic Design (ISQED). Santa Clara, CA, USA: IEEE; 2016

[17] Yegnanarayana B. Artificial Neural Networks. New Delhi, India: PHI Learning Pvt. Ltd.; 2009

[18] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;**521**(7553):436

[19] Koch C, Segev I. The role of single neurons in information processing. Nature Neuroscience. 2000;**3**(11s):1171

[20] Jaeger H. Echo state network. Scholarpedia. 2007;**2**(9):2330

[21] Lin X, Yang Z, Song Y. Intelligent stock trading system based on improved technical analysis and Echo State Network. Expert Systems with Applications. 2011;**38**(9):11347-11354

[22] Skowronski MD, Harris JG. Automatic speech recognition using a predictive echo state network classifier. Neural Networks. 2007;**20**(3):414-423

[23] Wang Q, Li Y, Li P. Liquid state machine based pattern recognition on FPGA with firing-activity dependent power gating and approximate computing. In: 2016 IEEE International Symposium on Circuits and Systems (ISCAS). Montreal, Canada: IEEE; 2016

[24] Zhang Y et al. A digital liquid state machine with biologically inspired learning and its application to speech recognition. IEEE Transactions on Neural Networks and Learning Systems. 2015;**26**(11):2635-2649

[25] Legenstein R, Maass W. Edge of chaos and prediction of computational performance for neural circuit models. Neural Networks. 2007;**20**(3):323-334

[26] Appeltant L et al. Information processing using a single dynamical node as complex system. Nature Communications. 2011;**2**:468

[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**(4):500-544

[28] FitzHugh R. Impulses and physiological states in theoretical models of nerve membrane. Biophysical Journal. 1961;**1**(6):445-466

[29] Abbott LF. Lapicque's introduction of the integrate-and-fire model neuron (1907). Brain Research Bulletin. 1999;**50**(5-6):303-304

[30] Liu Y-H, Wang X-J. Spike-frequency adaptation of a generalized leaky integrate-and-fire model neuron. Journal of Computational Neuroscience. 2001;**10**(1):25-45

[31] Zhao C et al. Neuromorphic encoding system design with chaos based CMOS analog neuron. In: 2015 IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA). Verona, NY, USA: IEEE; 2015

[32] Zhao C et al. Interspike-interval-based analog spike-time-dependent encoder for neuromorphic processors. IEEE Transactions on Very Large Scale Integration (VLSI) Systems. 2017;**25**(8):2193-2205

[33] Spruston N. Pyramidal neurons: Dendritic structure and synaptic integration. Nature Reviews Neuroscience. 2008;**9**(3):206

[34] Gerstner W, Kistler WM. Spiking Neuron Models: Single Neurons, Populations, Plasticity. Cambridge, United Kingdom: Cambridge University Press; 2002

[35] Zhao C et al. Spike-time-dependent encoding for neuromorphic processors. ACM Journal on Emerging Technologies in Computing Systems (JETC). 2015;**12**(3):23

[36] Panzeri S et al. Sensory neural codes using multiplexed temporal scales. Trends in Neurosciences. 2010;**33**(3):111-120

[37] Boumans T et al. Neural representation of spectral and temporal features of song in the auditory forebrain of zebra finches as revealed by functional MRI. European Journal of Neuroscience. 2007;**26**(9):2613-2626

[38] Reich DS et al. Interspike intervals, receptive fields, and

information encoding in primary visual cortex. Journal of Neuroscience. 2000;**20**(5):1964-1974

[39] Brasselet R et al. Neurons with stereotyped and rapid responses provide a reference frame for relative temporal coding in primate auditory cortex. Journal of Neuroscience. 2012;**32**(9):2998-3008

[40] Shao L et al. Spatio-temporal Laplacian pyramid coding for action recognition. IEEE Transactions on Cybernetics. 2014;**44**(6):817-827

[41] Zhao C et al. Energy efficient spiking temporal encoder design for neuromorphic computing systems. IEEE Transactions on Multi-Scale Computing Systems. 2016;**2**(4):265-276

[42] Zhao C, Li J, Yi Y. Making neural encoding robust and energy efficient: An advanced analog temporal encoder for brain-inspired computing systems. In: Proceedings of the 35th International Conference on Computer-Aided Design. Austin, TX, USA: ACM; 2016

[43] Bai K, Yi Y. A path to energy-efficient spiking delayed feedback reservoir computing system for brain-inspired neuromorphic processors. In: Proceedings of 19th International Symposium in Quality Electronic Design (ISQED). 2018

[44] Bai K, et al. Enabling an new era of brain-inspired computing: Energy-efficient spiking neural network with ring topology. In: 2018 55th ACM/ESDA/IEEE Design Automation Conference (DAC). San Francisco, CA, USA: IEEE; 2018

[45] Ehsan MA, et al. Design challenges and methodologies in 3D integration for neuromorphic computing systems. In: 2016 17th international symposium on Quality electronic design (ISQED). Santa Clara, CA, USA: IEEE; 2016

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

[47] An H, et al. Electrical modeling and analysis of 3D neuromorphic IC with monolithic inter-tier vias. In: 2016 IEEE 25th Conference on Electrical Performance of Electronic Packaging And Systems (EPEPS). San Diego, CA, USA: IEEE; 2016

[48] An H, et al. Electrical modeling and analysis of 3D synaptic array using vertical RRAM structure. In: 2017 18th International Symposium on Quality Electronic Design (ISQED). Santa Clara, CA, USA: IEEE; 2017

[49] 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 (EPEPS). San Jose, CA, USA: IEEE; 2013

[50] Ehsan MA, Zhou Z, Yi Y. Hybrid three-dimensional integrated circuits: A viable solution for high efficiency Neuromorphic Computing. In: 2017 International Symposium on VLSI Design, Automation and Test (VLSI-DAT). Hsinchu, Taiwan: IEEE; 2017

[51] Ehsan MA et al. A novel approach for using TSVs as membrane capacitance in neuromorphic 3-D IC. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems. 2018;**37**(8):1640-1653

[52] An H et al. Monolithic 3D neuromorphic computing system with hybrid CMOS and memristor-based synapses and neurons. Integration, the VLSI Journal. 2017:1-9. https://reader.elsevier.com/reader/sd/pii/S0167926017

303413?token=CEB3CFF7957972E96D2
8980DF318ADC0F8B4F025E8A9116BF9
E7CDA0B67D8C5F61CC84371F37CC1D5
4A5510D6206547F

[53] An H et al. Three dimensional memristor-based neuromorphic computing system and its application to cloud robotics. Computers and Electrical Engineering. 2017;**63**:99-113

[54] Wong H-SP et al. Metal–oxide RRAM. Proceedings of the IEEE. 2012;**100**(6):1951-1970

[55] 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). Pittsburgh, PA, USA: IEEE; 2017

[56] Clermidy F et al. Advanced technologies for brain-inspired computing. In: 2014 19th Asia and South Pacific Design Automation Conference (ASP-DAC). Suntec, Singapore: IEEE; 2014

[57] Swaminathan M. Electrical design and modeling challenges for 3D system integration. In: 12th International Design Conference, Dubrovnik, Croatia. 2012

[58] Gourier N, Hall D, Crowley JL. Estimating face orientation from robust detection of salient facial structures. In: FG Net Workshop on Visual Observation of Deictic Gestures. 2004