Engineering Nanoscale Devices for Brain-inspired Computing

Bipin Rajendran

Department of Electrical & Computer Engineering, New Jersey Institute of Technology, Newark, NJ, USA
**Corresponding author: bipin@njit.edu*

Abstract

There are several efforts worldwide to develop information processing platforms using nanoscale devices that mimic the key architectural features of the brain. In this paper, we review some of the recent efforts in engineering nanoscale devices that capture the dynamics of neurons and synapses in the brain for building neuromorphic computing systems. We will also discuss the challenges in building efficient systems based on these devices so that they can be ubiquitously deployed in energy-constrained environments such as IoT (Internet of Things) edge nodes and in enterprise computing for massively parallel data analytics and inference.

1. Introduction

Inspired by the organization and efficiency of the human brain, significant strides have been achieved in the field of machine learning over the past decade in developing algorithms capable of learning and inference on various complex cognitive tasks [1]. However, their implementation in von Neumann computing systems is highly non-optimal, due to the large data movement between the physically separated memory and processor units. This has motivated the search for post-CMOS materials and devices that could efficiently mimic the dynamics of the computational entities of these networks [2].

These networks consist of multiple layers of neurons connected through synapses whose weights are adjusted during training to minimize a cost function defined in terms of a labelled data-set. Such a network can be efficiently implemented in hardware using a tiled array of cross-bars [3], as illustrated in Figure 1. As opposed to deep learning networks, the third generation spiking neural networks (SNNs) use neuronal models that closely mimic the time based information encoding and processing aspects of the human brain – information is encoded in the time, rate or phases of issue of action potentials or spikes. Since spikes are issued sparsely, and all weight updates are triggered by spikes, SNN based hardware implementation could potentially be more energy efficient than the implementation of similar sized second generation ANNs. Though there are some recent efforts to build Si CMOS based hardware chips for deep learning [5], many noteworthy demonstrations rely on SNN architecture [6, 7], even though development of learning algorithms for them is a topic of active research today [8]. Hence, we will also focus on nanoscale devices for implementing neurons and synapses in SNNs in this paper.

2. Nanoscale Devices for Neuro-Synaptic Networks

The target specifications for nanoscale devices to implement neuronal and synaptic dynamics is listed in Figure 2. Clearly, achieving reliable operations in nanoscale devices at such low power is a significant challenge and requires integration and optimization of nanoscale materials and structures.

The high fan-out connectivity in these networks results in significantly larger number of synapses compared to the number of neurons. Hence, mimicking synaptic plasticity efficiently is the key objective of many hardware engineering efforts today. However, there are also recent demonstrations that aim to capture the integrate-and-fire dynamics of biological neurons based on phase transition in correlated oxides or chalcogenides [9,10].

Most of the synaptic engineering efforts are directed towards building devices that are capable of capturing several forms of spike-triggered conductivity modulation schemes observed in biological synapses [11] (Figure 3). One efficient way to implement STDP rules in memristive devices is to use programming waveforms that mimic the shape of action potentials that are issued by spiking neurons along both its input and output terminals – the amplitudes are chosen such that only the coincidence of waveforms from both the partner neurons of a synapse can alter its conductance [12] (Figure 4, 5). There are several demonstrations of synaptic plasticity in nanoscale devices based on variants of this technique, approaching sub-pJ levels [13] (Figure 6).

3. Discussion & Future Outlook

While individual devices have been demonstrated mimicking neuronal and synaptic dynamics, the joint co-optimization of learning algorithms, device characteristics and system operating profiles will be essential to meet the promise of computing systems that approach the efficiency of the brain.

References

[1] Y. LeCun et al., Nature. 521 (2015), 436-444.
[2] B. Rajendran et al., IEEE JETCAS (2016), 198-211.
[3] B. Rajendran et al., IEEE TED (2013), 246-253.
[4] N. Rodriguez et al., IEEE TED 56 (2009), 1507-1515.
[5] Y-H Chen et al., Proceedings of ISCA 2016.
[6] P. Merolla et al., Science (2014), 668-673.
[7] N. Qiao et al, Frontiers in Neurosc. (2015), 9.141.
[8] S. Kulkarni et al., Proc of IEEE MWSCAS, 2017.
[9] K. Moon et al., IEDM Tech Digest, 2015.
[10] T.Tuma et al., Nature Nanotechnology, 693–600, 2016.
[11] L. Abbot et al., Nature Neuro., 1178-1183, 2000.
[12] N. Panwar et al., Device Research Conference, 2014.
[13] S. Mandal et al., Nature Sci Reports 4:5333, 2014.

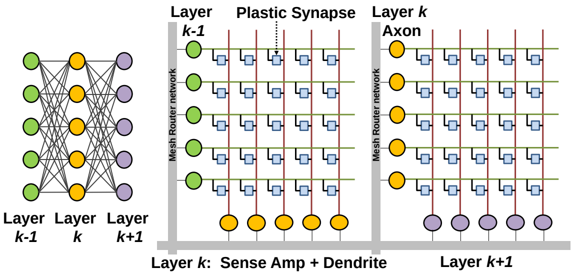
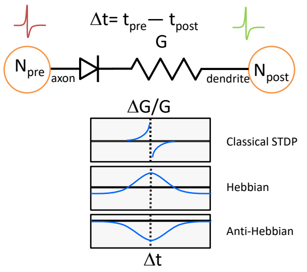
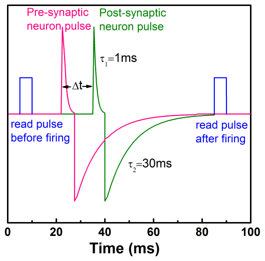


Fig. 1: Illustration of an artificial neural network and its equivalent cross-bar array based hardware implementation. Neuronal devices are in the periphery and synaptic devices are at the junctions of the cross-bar array. Arbitrary networks can be mapped into tiled arrays of cross-bars that communicates through binary spikes routed through a mesh network.

Attribute	Specification
Synapse	
Main characteristic	Analog programmability
Switching power, timescale	100 nW, 100 ns
Conductance Resolution	> 64 levels
Dynamic Range	~ 10 MΩ-100 MΩ
Programming non-linearity	$\Delta g(Vs)/\Delta g(Vs/2) > 0.1$
Area	$25 - 100\,F^2$
Switching endurance	$> 10^9$ programming cycles
Neuron	
Main characteristic	Intrinsic current integrating and thresholding behavior
Switching power, timescale	10 nW, 100 ns
Area	$< 10,000\,F^2$
Switching endurance	$> 10^{12}$ spikes

Fig. 2: Target specifications of nanoscale devices for brain-inspired computing systems.





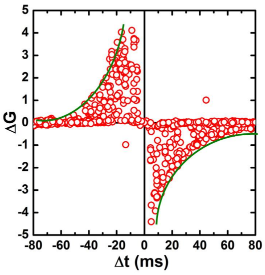


Fig. 3: Spike time dependent plasticity (STDP) is the modulation of effective conductivity ($\Delta G/G$) of the synapse based on the spike activity of pre- and post-synaptic neurons. In the equivalent circuit model, synaptic conductance G is modulated by the co-incident arrival of pre-and post-synaptic spikes. Three forms of plasticity rules are illustrated.

Fig. 4: Waveform engineering to implement spike timing dependent plasticity in a memristive device, adapted from [12]. The peak amplitudes of the pre and post-synaptic waveform is chosen such that they are below the minimum voltage required for perturbing the state of the device. Device conductivity modulation is hence a function of Δt.

Fig. 5: Experimental measurement of conductivity modulation in a PCMO thin film device mimicking biologically observed spike timing dependent plasticity based on the programming scheme in Figure 5, adapted from [12]. Arbitrary forms of plasticity can be obtained in memristive devices based on this waveform engineering technique.

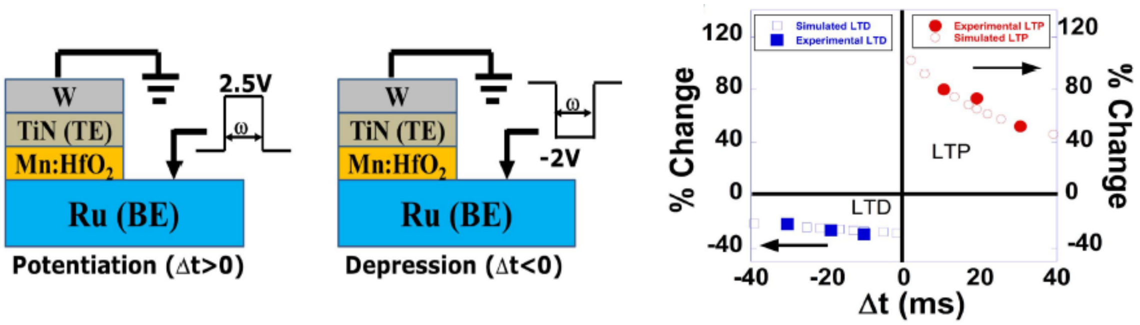


Fig. 6: Experimental demonstration of spike timing dependent plasticity in a TiN/Mn:HfO$_2$/Ru device based on a pulse-width modulation scheme, adapted from [13]. The energy consumed per event for this device is less than 500 fJ.