Peer-reviewed version available at J. Low Power Electron. Appl. 2018, 8, 34; doi:10.3390/jlpea8040034

Article

Towards Neuromorphic Learning Machines using Emerging Memory Devices with Brain-like Energy Efficiency

Vishal Saxena [1]*, Xinyu Wu [2], Ira Srivastava and Kehan Zhu [3]

[1] Electrical and Computer Engineering Department, University of Idaho; vsaxena@uidaho.com
[2] X. Wu was with University of Idaho, he is now with Micron Technology, Boise, ID, USA; tomas.wu@gmail.com
[3] K. Zhu was with Boise State University, he is now with Maxim Integrated, Beaverton, OR, USA; kehan.zhu@gmail.com
* Correspondence: vsaxena@uidaho.edu; Tel.: +1-208-885-6870

Abstract: The ongoing revolution in Deep Learning is redefining the nature of computing that is driven by the increasing amount of pattern classification and cognitive tasks. Specialized digital hardware for deep learning still holds its predominance due to the flexibility offered by the software implementation and maturity of algorithms. However, it is being increasingly desired that cognitive computing occurs at the edge, i.e. on hand-held devices that are energy constrained, which is energy prohibitive when employing digital von Neumann architectures. Recent explorations in digital neuromorphic hardware have shown promise, but offer low neurosynaptic density needed for scaling to applications such as intelligent cognitive assistants (ICA). Large-scale integration of nanoscale emerging memory devices with Complementary Metal Oxide Semiconductor (CMOS) mixed-signal integrated circuits can herald a new generation of Neuromorphic computers that will transcend the von Neumann bottleneck for cognitive computing tasks. Such hybrid *Neuromorphic System-on-a-chip (NeuSoC)* architectures promise machine learning capability at chip-scale form factor, and several orders of magnitude improvement in energy efficiency. Practical demonstration of such architectures has been limited as performance of emerging memory devices falls short of the expected behavior from the idealized memristor-based analog synapses, or weights, and novel machine learning algorithms are needed to take advantage of the device behavior. In this work, we review the challenges involved and present a pathway to realize ultra-low-power mixed-signal NeuSoC, from device arrays and circuits to spike-based deep learning algorithms, with 'brain-like' energy-efficiency.

Keywords: Cognitive Computing, Deep Learning, Neuromorphic System-on-a-Chip (NeuSoC), NVRAM, RRAM, Silicon Neurons, Spiking Neural Networks (SNNs).

1. Introduction

Recent grand challenge in semiconductor technology urge researchers to "Create a new type of computer that can proactively interpret and learn from data, solve unfamiliar problems using what it has learned, and operate with the energy efficiency of the human brain [1]." Artificial Intelligence (AI) techniques such as deep neural networks, or deep learning, have found widespread success when applied to several problems including image and video interpretation, speech and natural language processing, and medical diagnostics [2]. At present, much of cognitive computing is performed on digital graphical processing units (GPUs), accelerator ASICs, or FPGAs, mostly at the data center end of the Cloud infrastructure. However, the current explosion in widespread deployment of deep-learning applications is expected to hit a power-performance wall with-(1) plateauing in CMOS scaling, and (2) limits set for energy consumption in the Cloud. These deep learning implementations

Peer-reviewed version available at J. Low Power Electron. Appl. 2018, 8, 34; doi:10.3390/jlpea8040034

take long computing cluster days to train a network for realistic applications. Even with the remarkable progress made in computing, the nimble human brain provides an existential proof that learning can be more sophisticated while allowing compactness and energy-efficiency. Furthermore, there is a growing interest in edge computing and intelligent cognitive assistants (ICAs), where deep learning and/or inference are available on energy-constrained mobile platforms, autonomous drones, and internet-of-things sensor nodes, which not only eliminate the reliance on cloud-based AI service, but also ensure privacy of user data.

In contrast to the predominant von Neumann computers where memory and computing elements are separated, a biological brain retains memories and performs 'computing' using largely homogeneous neural motifs. In a brain, neurons perform computation by propagating spikes and storing memories in the relative strengths of the synapses, and by forming new connections (or morphogenesis). By repeating these simple cortical columnar organization of neurons and synapses, a biological brain realizes a highly energy-efficient cognitive computing motif. Inspired by biological nervous systems, artificial neural networks (ANNs) were developed which have achieved remarkable success in a few specific applications. In the past decade, by leveraging parallel graphics processing units (GPUs), ASICs [3], or field-programmable gate arrays (FPGAs), power consumption of artificial neural networks has been reduced but yet remains significantly higher than their biological counterpart, developed through millions of years of evolution. The discovery of spike-timing-dependent- plasticity (STDP) local learning rule [4,5] and mathematical analysis of spike-based winner-take-all (WTA) motifs have opened new avenues in spike-based neural network research. Recent studies have suggested that STDP, and its neural-inspired variants, can be used to train spiking neural networks (SNNs) in-situ without trading-off their parallelism [6,7].

In this work, we present architectural overview, challenges associated with the interplay of emerging non-volatile memory devices, circuits, and algorithms and their mitigation for practical realization of NeuSoCs. This paper is organized as follows. Section 2 presents an overview of existing neuromorphic computing platforms and the potential for nanoscale emerging memory devices. Section 3 presents a review on the mixed-signal approach to neuromorphic computing leveraging crossbar arrays of emerging memory devices and details on neural circuits and learning algorithms, followed by challenges associated with emerging devices. Section 4 makes an argument for bioplausible dendritic computing using compound stochastic synapses. Section 5 discusses energy-efficiency implications of device properties on neuromorphic SoCs. Section 6 presents the direction for algorithm development for large scale deep learning using neurmorphic substrates, followed by conclusion.

2. Neuromorphic Computing and Emerging Devices

2.1. Digital Neuromorphic Platforms

Recent progress in neuromorphic hardware has led to development of asynchronous event-driven, as opposed to synchronous or clock-driven, ICs that process and communicate information using spatio-temporal voltage spike signals. Most pertinent example of a digital neuromorphic hardware are IBM's TrueNorth [8], SpiNNaker system from the European Brain Project [9], and recently Loihi chip from Intel [10]. IBM's TrueNorth ASIC comprises of 4096 cores, with 1 million programmable neurons and 256 million programmable synapses as communication channels between the digital neurons, and consumes $\approx 100mW$ for pattern classification tasks [8]. However, the networks are trained offline as the chip doesnt allow in-situ learning. On the other hand, Intel's Loihi ASIC implements on-chip learning with flexibility in neuron and synpase behavior, but trades off learning with neurosynaptic density [10]. Purely digital implementations have low neurosynaptic density and large die area which can limit the scalability and cost of the resulting neuromorphic systems. Further, leakage power in SRAM-based digital synapses limits the overall energy-efficiency.

2.2. Subthreshold Analog Neuromorphic Platforms

Advances in analog neuromorphic circuits include the Neurogrid hardware [11], where subthreshold biomimetic CMOS circuits are developed to reproduce dynamics occurring in biological neural networks. These implementations leverage the fact that the brain performs analog-like spike-based computation with a massive number of imprecise components. However, the fundamental limitation of such architectures is that the weights are dynamically stored and updated on capacitors, which leak away in few milliseconds, limiting any long-term learning. Bistability of analog weights has been used as an intermittent solution [12]; however, recent studies on deep neural networks have determined that at least 4-bit resolution is needed for the synaptic weights to realize meaningful learning system [13].

2.3. Neuromorphic Platforms using Floating-gate and Phase Change NVMs

Other solutions include using floating gate, or Flash memory, devices [14,15] and phase change memory (PCM) [16,17] for implementing non-volatile synaptic weights. The endurance of floating-gate devices is typically $\approx 100k - 500k$ cycles due to the high voltages (5-18V) used for program and erase. This will preclude on-chip training of neural networks where millions of program/erase operations are anticipated. Flash memory is best suited for low-power inference applications, or for scenarios where learning concludes within the endurance limit of the devices.

Recently, IBM's neuromorphic group has shown encouraging results in the use of PCM devices from use as synapses in SNNs [16,18]. PCM devices can provide incremental states in the program direction by controlling the amount of crystallization on the memory cell. However, the erases can be abrupt as the device undergoes a melt-and-quench phase when brought to the amorphous state [16].

2.4. Nanoscale Emerging Devices

In the last decade, there has been a renewed interest in two-terminal resistive memory devices, including the elusive memristor, as these resistive random access memory (RRAM) and similar devices promise very high density ($Terabits/cm^2$) [19]. These devices have demonstrated biologically plausible STDP plasticity behavior in several experiments [19,20], and therefore have emerged as an ideal candidate for realizing electronic equivalent of synapses. Also, recent advances in these devices have shown low-energy consumption to change their states with sub-100fJ switching energy and very compact layout footprint ($F = 10nm$ pitch with $4F^2$ cell size) [21–23]. Following this trend, hybrid CMOS-RRAM analog very-large-scale integrated (VLSI) circuits have been proposed [24,25] to achieve dense integration of CMOS neurons with these emerging devices for neuromorphic computing chips by leveraging the contemporary nanometer silicon processing technology.

The author also introduced a first compact CMOS memristor emulator circuit [26,27] and the resulting dynamic synapse circuits [28] but concluded that non-volatile synapses are needed for long-term retention of weights, high synaptic density, and low leakage power in trained neural networks. Consequently, the Neuromorphic computing architecture development requires synergistic development in devices, circuits and learning algorithms to take advantage of the high synaptic density while not being oblivious to the challenges at the device-circuit interface. Following four necessary criterion have been identified for realizing large scale NeuSoCs capable of deep learning:

1. Non-volatility and high-resolution of the synaptic weights,
2. High neurosynaptic density, approaching billions of synapses and millions of neurons per chip,
3. Massively-parallel learning algorithms with localized updates (or in-memory computing)
4. Event-driven ultra-low-power neural computation and communication.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 September 2018 doi:10.20944/preprints201807.0362.v2

Peer-reviewed version available at J. Low Power Electron. Appl. 2018, 8, 34; doi:10.3390/jlpea8040034

3. Mixed-Signal Neuromorphic Architecture

Mixed-signal neuromorphic ICs promise the potential for embedded learning and pattern classification with orders of magnitude lower energy consumption than the von Neumann processors. As discussed in the previous sections, this is feasible due to the densely-integrated non-volatile memory devices that include the resistive random access memory (RRAM) [29,30], phase-change memory, conductive-bridge RAM (CBRAM) [31], STTRAM [32], and 3D crosspoint memory [33]. These are also referred as memristors or memristive devices in literature [19,34].

3.1. Crossbar Networks

CMOS neurons and RRAM synapses are organized in a crossbar network to realize a single-level of neural interconnections as shown in Figure 1 [25,35]. In this architecture, each input neuron is connected to another output neuron through a two-terminal RRAM to form a crossbar, or cross-point array. By cascading and/or stacking such crossbars, a deep neural network can be realized in hardware. Further, maximum synaptic density is achieved by minimizing or eliminating the overheads associated with the synapse, while transferring the complexity to the peripheral neurons, as opposed to random access memory architectures. The crossbar architecture is tolerant to sneak-paths in the array as all devices are concurrently used in the neural network, as opposed to the random access case where RRAM bit(s) are accessed and read out. Consequently, the sneak paths are absorbed into the network weights without significant performance degradation. Further, advanced packaging techniques such as through silicon via (TSV) for multiple chips and flip-chip integration can be leveraged to realize 3D stacking of such networks.

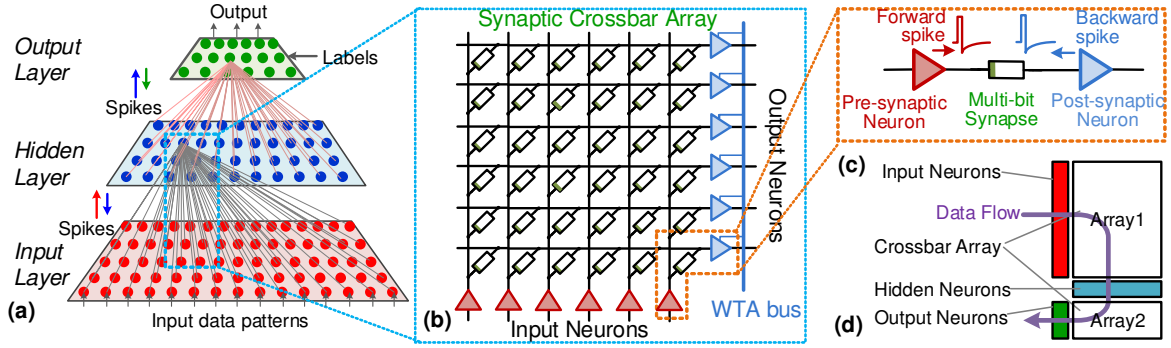

Figure 1. Neuromorphic SoC architecture: (a) A fully-connected spiking neural network showing input, hidden and output layers comprised of spiking neurons. Here, synaptic connections shown for one neuron in the hidden and output layers; (b) A slice of the neural network architecture implemented using RRAM crossbar memory array and column/rows of mixed-signal CMOS neurons with shared bus architecture for competitive learning; (c) A single multi-bit synapse between the input (pre-synaptic) and output (post-synaptic) neurons that adjusts its weight using STDP; (d) the architecture leverages 2D arrays and peripheral circuits used in memory technology to achieve high-density spiking neural network hardware.

3.2. Analog Synapses using RRAM/Memristors

Several nano-scale RRAM or memristors in literature have shown that their conductance modification characteristics are similar to the Spike-timing dependent plasticity (STDP) rule from neurobiology [23,36,37], and thus are potentially an ideal candidate for implementing electronic synapses. STDP states that the synaptic weight w is updated according to the relative timing of the pre- and post-synaptic neuron firing. This is a form of Hebbian learning that postulates that "neurons that fire together, wire together [38]." As illustrated in Figure 2 (a), a spike pair with the pre-synaptic spike arrives before the post-synaptic spike results in increasing the synaptic strength

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 September 2018 doi:10.20944/preprints201807.0362.v2

Peer-reviewed version available at J. Low Power Electron. Appl. 2018, 8, 34; doi:10.3390/jlpea8040034

(or long-term potentiation, LTP); a pre-synaptic spike after a post-synaptic spike results in decreasing the synaptic strength (or long-term depression LTD). Changes in the synaptic weight plotted as a function of the relative arrival timing of the post-synaptic spike with respect to the pre-synaptic spike is called the STDP learning function or learning window. Furthermore, during the inference mode, only the pre-spikes with the positive rectangular pulse are used for carrying the feedforward inputs through the SNN. The post-spikes and the negative tails are activated during the training mode only to enable on-chip learning. This not only saves energy but also avoids undesirable changes to the synaptic weights.

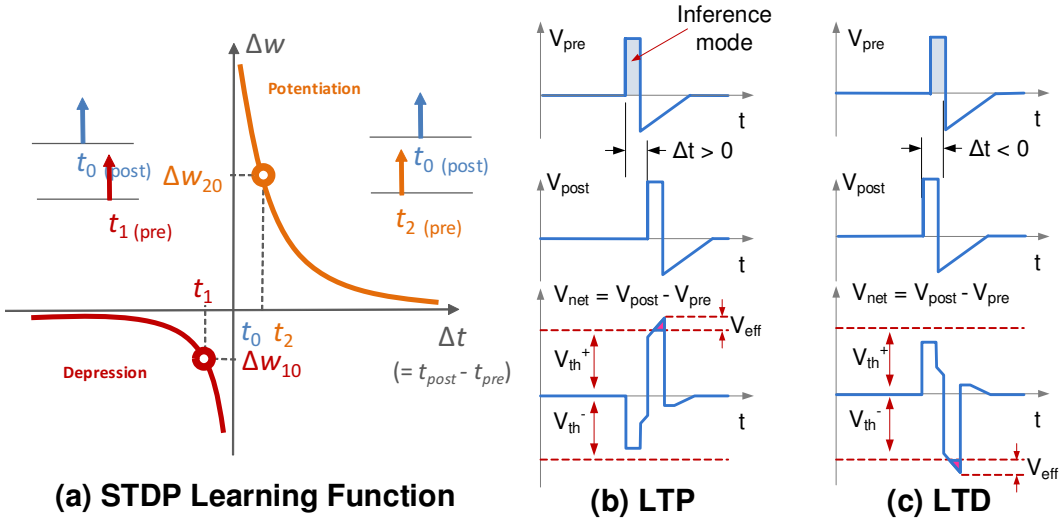

Figure 2. A two-layer spiking neural network crossbar RRAM synapse array and peripheral CMOS neurons with WTA bus. During the inference mode only the pre-spikes with the positive rectangular pulse are used; the post-spikes and the negative tails are activated during the training mode.

In pair-wise STDP learning, spikes sent from pre- and post-synaptic have their voltage amplitudes below the program and erase switching thresholds (V_{th}^{+} and V_{th}^{-}) of a bipolar RRAM device. RRAM switching events may occur only if this spike pair overlaps and creates a net potential (V_{net}) greater than the switching threshold, as illustrated in Figure 2 (b,c). This scheme effectively converts the time overlap of spikes into program or erase voltage pulses [39,40]. In case of no temporal overlap, the pre-synaptic pulse is integrated in the neuron and thus should have a net positive area and smaller amplitude than the program or erase thresholds.

3.3. Event-driven Neurons with Localized Learning

Driving thousands of resistive devices in parallel while maintaining desired energy-efficiency presents difficult challenges for the CMOS neurons. The authors earlier demonstrated low-power integrate-and-fire neuron circuits that can drive memristor/RRAM synapses with in-situ spike-timing dependent plasticity (STDP) based learning [39]. This is illustrated in Figure 3 where a single opamp-based design is employed so that the neuron can drive the resistive load presented by the RRAM synapses [25,39].

The neuron operates in four event-driven modes as shown in Figure 4. In the normal integrating mode during training or inference, they are biased with very low current ($< 1\mu A$) and integrate the incoming spikes weighted by the RRAM conductance ($i_k = \sum_j w_{kj} \cdot V_{spk,j}(t)$). When the integrated membrane potential, $V_{mem,j}$, crosses the threshold V_{thr}, a firing event occurs whereby the neuron is reconfigured as a voltage buffer and dynamically biased with large current so as to drive the RRAM synapses.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 September 2018 doi:10.20944/preprints201807.0362.v2

Peer-reviewed version available at J. Low Power Electron. Appl. 2018, 8, 34; doi:10.3390/jlpea8040034

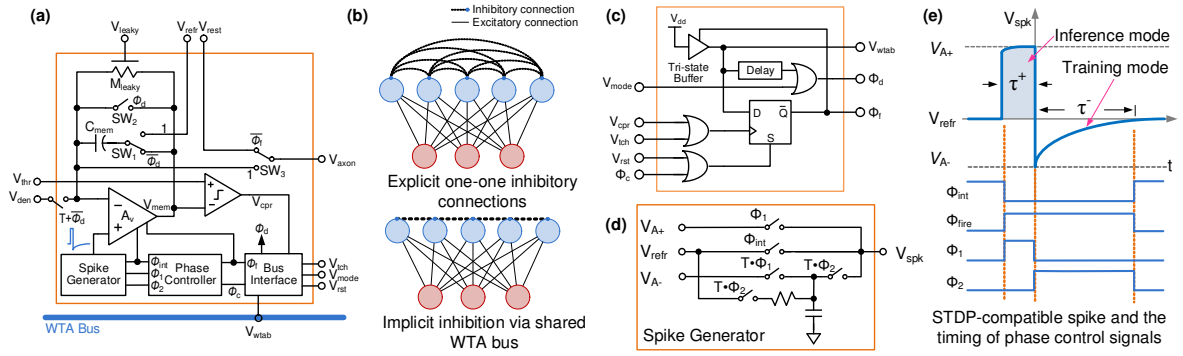

Figure 3. (a) Schematic of the integrate-and-fire Neuron for neural learning. (b) Competitive learning uses explicit one-on-one inhibitory connections, whereas the same function can be implemented with implicit inhibition on a shared WTA bus. (c) The proposed asynchronous WTA bus interface circuit. (d) Spike generator circuit for spikes with rectangular positive tail during the training as well as inference mode, and an exponential negative tail during the training mode (T=1) [25].

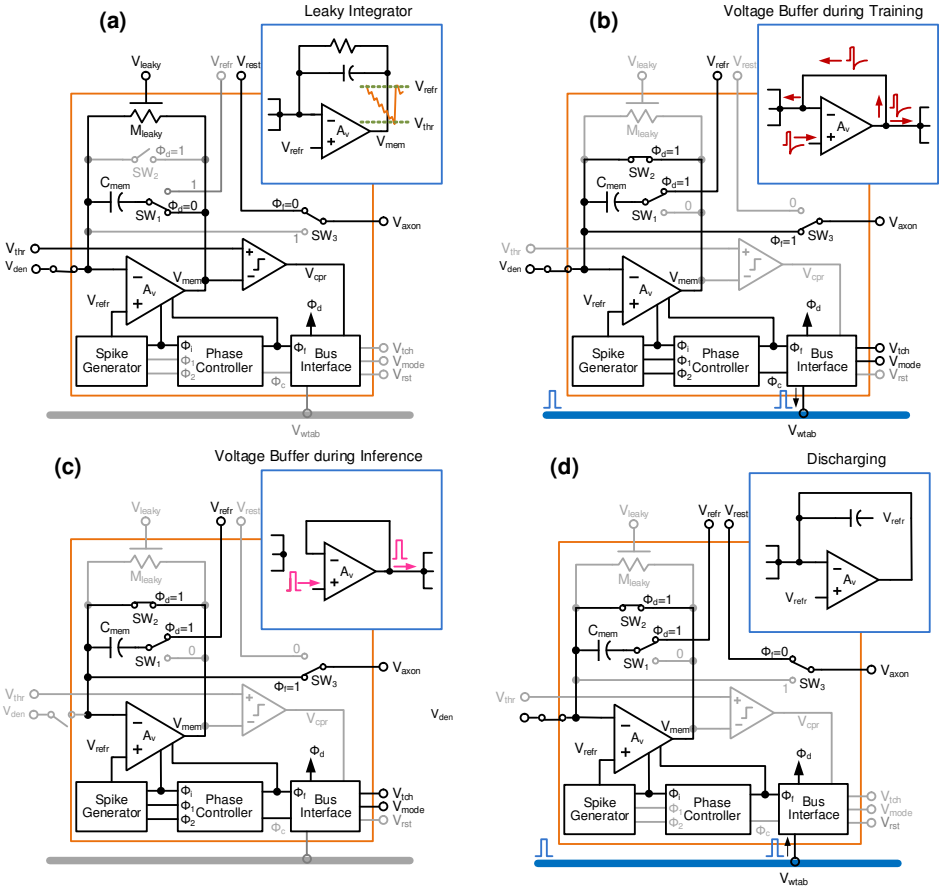

Figure 4. Event-driven operation of the proposed leaky integrate-and-fire neuron during training and inference.

During training, i.e. when the signal $T = 1$, the voltage spikes with positive pulse and negative tail are propagated in the forward (pre spikes) as well as the backward direction (post spikes). This enables learning by adjusting the synaptic weights (w_{kj}) using STDP based 'write' mechanism seen in Figure 10. During inference ($T = 0$), only the pre-spikes are propagated in the forward direction,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 September 2018 doi:10.20944/preprints201807.0362.v2

Peer-reviewed version available at J. Low Power Electron. Appl. 2018, 8, 34; doi:10.3390/jlpea8040034

and that too with the positive header. Here, no learning takes place and the synaptic weights are preserved while 'reading' them.

After the spike event concludes, the neuron returns to the background integrating mode after a refractory period τ_{refr}. A fourth mode, called discharge mode, allows competition between neurons. All the neurons are connected using a shared winner-take-all (WTA) bus; if a winner neuron fires first, other neurons are discharged to discourage them from spiking, forming a powerful neural learning motif [25]. A chip was designed using an earlier version of this neuron where associative learning (Pavlov's dog experiment) was demonstrated [39] as shown in Figure 5.

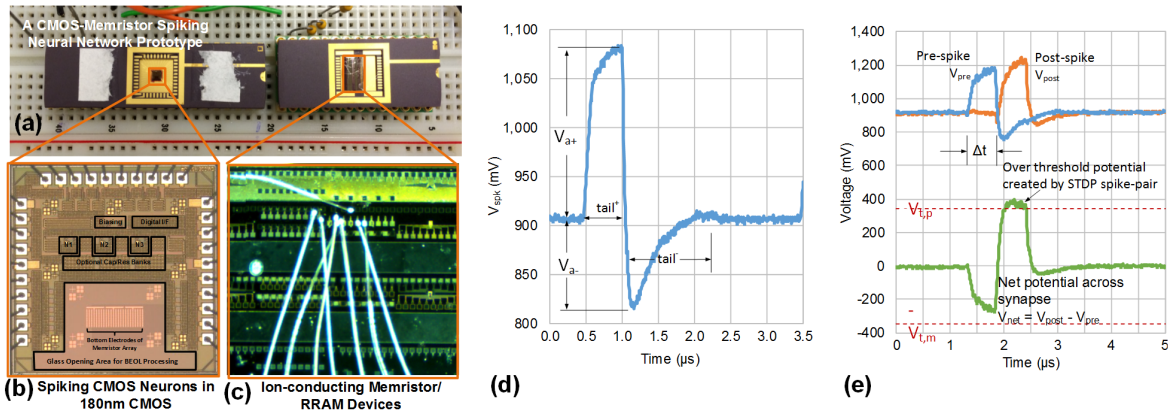

Figure 5. RRAM-compatible CMOS Neuron: (a) A CMOS-RRAM experimental prototype with (b) 180-nm CMOS spiking neuron chips with digital reconfigurability, and (c) possible interfacing with CBRAM devices [41]. (d) Measured spike output for one of the settings, (e) Pre- and post-spike voltage difference applied across a synapse [39].

3.4. Spike-based Neural Learning Algorithms

Spiking neural networks (SNNs) are gaining momentum due to their biological plausibility as well as the potential for low-power hardware implementation. Recently, it was analytically shown that the WTA with exponential STDP realizes a powerful unsupervised learning motif that implements expectation maximization; network weights converge to the log probability of the hidden input cause [7,42]. The authors developed algorithms that were compatible with the presented circuits to demonstrate general-purpose pattern recognition engine that consumes ultra-low energy, and were applied to handwritten digit recognition [25,40]. A winner-take-all (WTA) shared bus architecture, with novel event-driven switched-capacitor CMOS neurons, was demonstrated that allows unsupervised as well as supervised competitive learning with significant reduction in hardware complexity and chip layout area [25]. This two-layer network was simulated with transistor-level circuits using Cadence Spectre for the UCI 8×8 handwritten digit recognition task. Here, a teacher signal is used that only allows the desired neuron to fire, based on WTA+STDP, for a given output label in the training set.

This semi-supervised spiking network achieved a classification accuracy of 94% for four digits and 83% on all ten digits for with around 1000 training samples for each image label. Here, Figure 6 shows the evolution of synaptic weights for each of the ten output neurons as the learning progresses during the training period. Here, we can see that each neuron specializes in detecting only one of the digits and multilevel weights allow higher classification accuracy by emphasizing on critical features of the digits. In experiments with binary synapse models, the classification accuracy drops below 80%.

Peer-reviewed version available at J. Low Power Electron. Appl. 2018, 8, 34; doi:10.3390/jlpea8040034

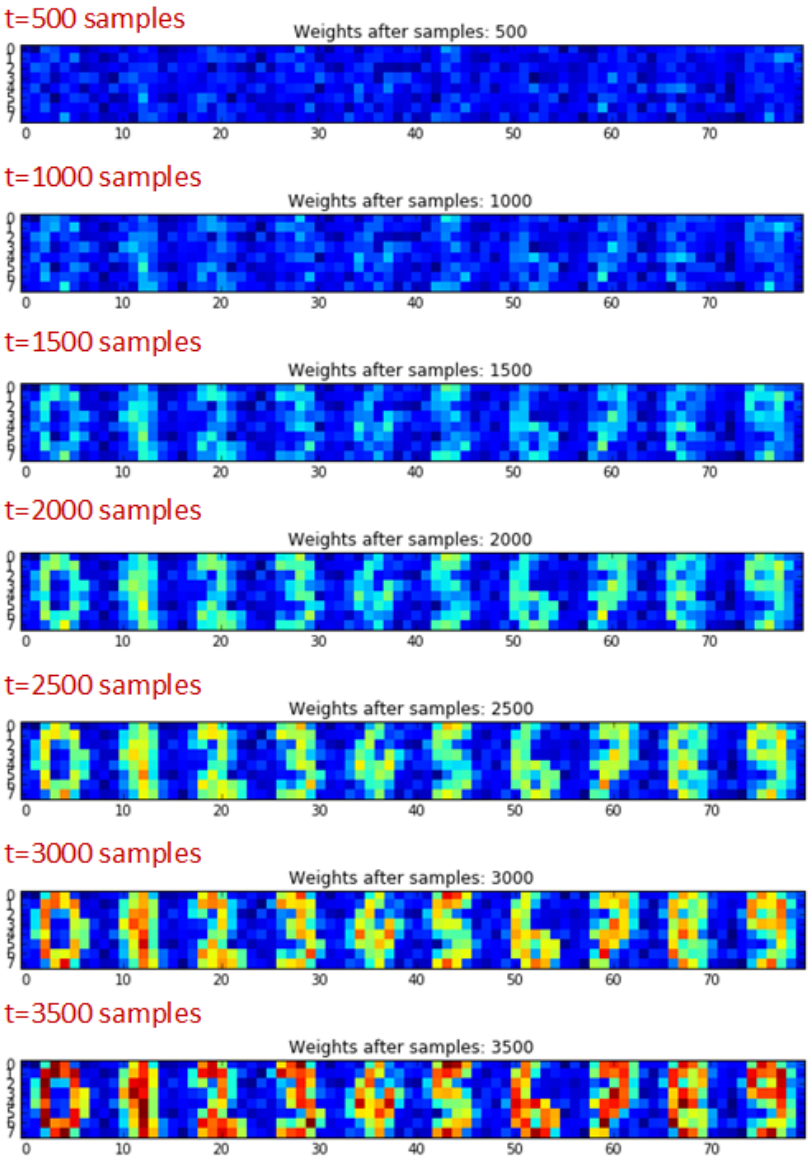

Figure 6. Evolution of simulated synaptic weights (normalized to the color scale) in the SNN for 8×8 handwritten character classification.

Higher classification accuracy can be potentially achieved by increasing the number of competing neurons [43] and/or stacking these spiking WTA motifs with backpropagation (backprop) algorithm adapted to the SNNs; a challenging task due to the non-differentiable nature of the spiking neurons. Recently, there was a successful demonstration of transfer learning whereby first a standard deep ANN was trained and its weights were then transferred to an equivalent SNN achieving close to 99% accuracy [44], followed by a spiking backprop that used membrane potential as a differentiable function [45]. In parallel, semi-supervised deep spike-based convolutional networks (ConvNets) for image pattern classification using GPUs have claimed >98% classification accuracy [46,47]. Moreover, there is a growing interest in developing backprop for deep SNNs with some success [48]. Even though spike-based backprop, in its current form, may not be the actual algorithm responsible for computation occurring in a biological brain. Nevertheless, it provides an intermittent solution to cognitive applications desired by the computing community. Needless to say, development of learning algorithms for SNN is a promising area of research and in conjunction with the field of computational neuroscience may lead to better understanding of brain computation. However, these

algorithms must be re-casted based upon the behavior of the synaptic devices such as in [17], where the STDP was modified to accommodate abrupt reset (depression) in PCM-based synapses.

3.5. Challenges with Emerging Devices as Synapses

Contemporary memristive or RRAM devices exhibit several limitations when considered for realizing neuromorphic computing:

(1) Resistive Loading: Resistive loads are typically avoided in CMOS circuits due to the resulting static power consumption. Consequently, large load resistance range is desirable to minimize power consumption in the CMOS neuron circuits that would drive a large number of such resistive devices in parallel. Ideally a value of $10 - 100M\Omega$ for the low resistance state (LRS) or 'On' state is needed for low power consumption. Further, very large LRS resistance, say $1G\Omega$, will result in extremely low signal-to-noise ratio (SNR) in presence of circuit noise. Thus, device geometry and the material stack needs to take these constraints into consideration.

(2) Variability and Stochasticity: RRAM devices exhibit significant variations (across different devices) and stochasticity (in the same device) in their behavior. This is observed as the program/erase threshold voltages ($V_{th^{+/-}}$) exhibit stochasticity and variability that in turn depends upon: (1) the initial 'electroforming' or 'breaking-in' step where the filament is formed in a pristine RRAM cell [49]. The program threshold voltage required for creating a filament (or phase change in the bulk) depends upon the compliance current (I_{CC}) and consequently the range of resistance for the LRS state. For example, a low compliance of $I_{CC} = 50nA$ creates a narrow and weak filament which displays analog-like incremental resistance change behavior, concomitant with large variation in the LRS resistance. What's more is that these analog-like states may relax to another value within seconds to hours. On the other hand, a large compliance current, say $I_{CC} = 5\mu A$, results in a thick filament that exhibits bistable switching behavior with lower variance in the LRS range. Moreover, independently setting the compliance current in a crossbar array in a NeuSoC is unwieldy due to large circuit overhead incurred.

(3) Resolution and Retention: Experimental studies have shown that it can be challenging to obtain stable weights for more than a single-bit resolution in RRAMs, especially without applying compliance current. In some studies, multi-level resistance in oxide-based memristive devices has been observed by fine tuning the device fabrication and/or electrical pulses for program and erase [49,50]. The analog state retention in actual crossbar circuit configuration is presently being studied [50]. Fundamentally, realizing multiple stable resistance states can be challenging due to the fact that a sufficiently large energy barrier is needed to separate two adjacent resistance states, which is not overcome by thermal energy, leakage or disturbance during the read (inference) operation.

Furthermore, we we recently showed using a simple CMOS emulator circuit that the pinched hysteresis characteristics of a conceptual memristor doesn't guarantee analog retention [27,28]. Based on this discussion, we can assume the worst case scenario that many such RRAM devices in crossbar arrays, without setting compliance current, may end up as bistable nonvolatile memory cells.

(4) Polarity: Several RRAM and CBRAM devices exhibit unipolar switching (i.e. the erase threshold V_{th^-} is much smaller in magnitude than the program threshold V_{th^+}). This is due to the fact that only a small amount of negative voltage can break or dissolve the filament and erase the device to its high resistance state (HRS). Such unipolar switching characteristics may not be compatible with the STDP scheme shown earlier in Figure 3 and may require circuit modification at the neuron-synapse interface.

(5) Endurance: Since training algorithms continually update network weights while being trained on massive amount of data, synptic device endurance ultimately governs the in-situ on-chip learning capability in a NeuSoC chip. For example, floating-gate or Flash devices are rather better suited for inference tasks due to $\approx 10^5$ cycles of write endurance, while phase-change memory devices last for $< 10^8$ write cycles. On the other hand RRAM devices are expected to endure more than 10^9 write cycles which makes them promising for continuously learning on a chip.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 September 2018 doi:10.20944/preprints201807.0362.v2

Peer-reviewed version available at J. Low Power Electron. Appl. 2018, 8, 34; doi:10.3390/jlpea8040034

4. Bio-inspiration for Higher-resolution Synapses

Presynaptic and postsynaptic neural activity enables the chemical synapses to change their weights or strengths of connection via biological mechanisms such as long-term depression (LTD) and long-term potentiation (LTP) in an adult human brain. This activity dependent synaptic plasticity is evidently the basis of learning and memory in human brain. As the evidence of role played by activity dependent synaptic plasticity in learning and memory gathers, our understanding of the underlying 'algorithm' for cognition in the brain also evolves.

As shown in Figure 7, a biological neuron cell has a body called 'soma' with a long axonal tail. The axon branches into axonal terminals or 'telodendria.' The soma has projections called dendrites. Synapses are created at the junction between axon terminal of presynaptic neuron and the dendrite on the postsynaptic neuron (soma). Each axonal terminal comprises of thousands of synaptic vesicles, which in turn each contain thousands of neurotransmitter molecules. Neurotransmitters are biochemical molecules that play vital role in signal transduction between the neurons. In response to an electrical stimulation and resulting Ca^{2+} influx in the pre-synaptic axon terminal, neurotransmitters are released from synaptic vesicles where they are stored into the synaptic cleft. These biochemical molecules then bind to their specific receptors in the dendrites of the post-synaptic neuron which eventually lead to opening of ligand-gated ion channels and thus generating an action potential. The whole process takes under two milliseconds of time [51]. The timing between presynaptic and postsynaptic action potential determines the synaptic plasticity and is mediated through biological events such as long term potentiation and depression of synaptic transmission. The action potential which travels across the axon of postsynaptic neuron is also responsible for initiating a voltage spike in dendrites from which it originated known as backpropagating action potential (spike). This is now known to be a critical step in synaptic plasticity and involves calcium influx into the dendritic spine.

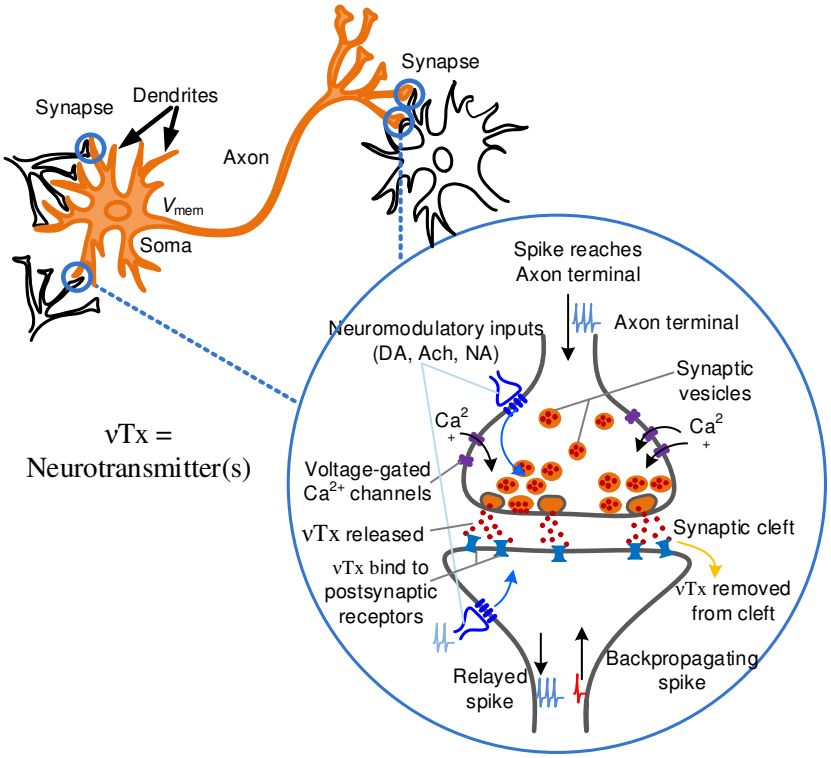

Figure 7. Synapses are present at the junction of axonal terminal and dendrites of the biological neurons. Some of the known neurotransmitter (νTx) signaling mechanisms are illustrated.

The timing of the spike, amount of calcium influx and distance of dendrites from neuron body determines the degree of the LTP. Dendrites also play a role in neural signal processing through signal attenuation and potentially modification of STDP. The LTD mechanism is still not yet well established. The signal transduction pathways activated by calcium transients in dendrites can impact plasticity by gene activation and synthesis of new proteins which can further modify neurotransmitter release from presynaptic membrane and the number of receptors at the post synaptic membrane. These biological processes are known to affect synaptic plasticity and has been a source of continued research and insight for understanding the learning rules at the individual synaptic level. An axon terminal can contain more than one type of neurotransmitter. The small molecule neurotransmitters such as glutamate, glycine mediate the fast responses of synaptic transmission whereas the neuropeptide transmitters, as well as the biogenic amines like (DA) and acetylcholine (Ach) and some other small-molecule neurotransmitters, are involved in regulation of neuronal activity and thus the learning rate [51]. Thus the principle neuromodulatory effect is to gate plasticity by modifying the spike-timing-dependent plasticity (STDP) learning window [52].

Further understanding of neuromodulation mechanisms will help us determine the actual learning mechanism in the brain at the abstraction level of large networks. We now know that dendrites also have a role through nonlinear spike processing and potential modification of STDP [53]. From a signal processing perspective, we can think of the neuromodulators release phenomenon as a quantized and stochastic mechanism. Consequently, plasticity can also be thought of as discrete and stochastic in nature, instead of continuous and analog. However, it has been experimentally difficult to verify this hypothesis or its contrary. Thus, if the biological synapses were binary, average over thousands of synapses combined with nonlinear dendrites will render their short and long-term plasticity to be analog and continuous in neurobiology experiments.

There is a continual flow of ideas from the computational neuroscience community where they mathematically model and analyze the underlying principles behind neural computing and the role of plasticity, neuromodulation and inhibition. Novel insights lead to refinement of learning algorithms with an ultimate goal of replacing backpropagation by a more biology-like unsupervised and lifelong learning. Implementation of these ideas in circuits follows naturally. At this point, the role of supporting neural cells such as glia and astrocytes, which comprise of almost half the neural mass, are not well understood; they are ignored in neuromorphic computing till their role becomes clear and significant.

4.1. Compound Synapse with Axonal and Dendritic Processing

The limitations of current memristive or RRAM devices pose a challenge to the realization of continuous-valued synaptic weights with reasonable resolution. Recent work has demonstrated binary-weighted spiking neural networks (SNNs) with 1% to 3% drop in classification accuracy [54]. Other SNN studies have established that synaptic weights with ≥ 4 bit resolution are required for no significant loss of accuracy [55]. In order to obtain more than binary resolution with the worst-case scenario of bistable RRAM devices, compound synapses were introduced in [56]. Here, several (say $M = 10$) stochastic memristors were employed in parallel to obtain an approximate resolution of $log_2 M$ bits on average. This concept was extended to include presynapstic axonal attenuation with parallel stochastic switching RRAMs [57]. Combination of such processing with parallel synapses allows additional degrees of freedom that enable fine-tuning of the STDP learning characteristics while emulating higher resolution plasticity [57].

In this work, the concept is extended to combine axonal (presynaptic) as well as dendritic (postsynaptic) processing. This is shown in Figure 8 where parallel RRAM devices are organized in a 2D sub-array structure with spike attenuation is introduced in pre as well as postsynaptic path. The postsynaptic path is analogous to dendritic processing in biology, while axonal processing is an artificial modification. Fundamentally in this configuration for the same pre and post-spike delay (Δt) each stochastic RRAM device sees different pulse voltages across it and thus has distinct probability

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 September 2018 doi:10.20944/preprints201807.0362.v2

Peer-reviewed version available at J. Low Power Electron. Appl. 2018, 8, 34; doi:10.3390/jlpea8040034

of switching. Thus, a smaller Δt causes more individual RRAMs to switch than a larger value of Δt, thus providing better control of the STDP behavior.

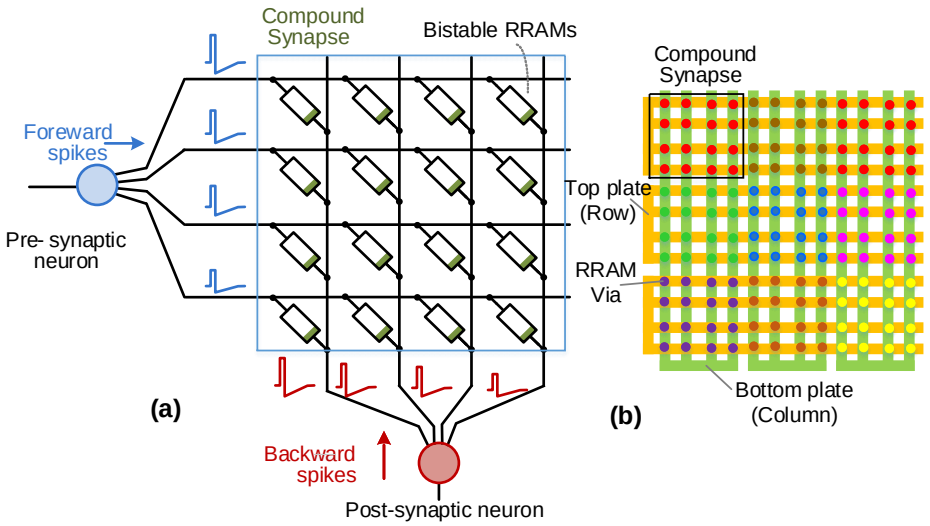

Figure 8. (a) A compound synapse in a 4×4 dendritic configuration with M=16 parallel bistable RRAMs, (b) a possible layout configuration for the compound synapse.

The proposed compound synapse concept with dendritic processing is biologically plausible as the STDP updates are discrete and stochastic, similar to the probabilistic release of the neurotransmitters. When averaged over a large number of synapses with individual dendritic attenuation, the discrete probabilistic plasticity emulates continuous analog-like behavior. Figure 9 shows the simulation results for the proposed concept. In this simulation, $M = 16$ RRAMs are employed with pre and post synaptic attenuations. Assuming Gaussian distribution of the program/erase threshold voltages, the stochastic switching behavior of the bistable RRAM device is given by cumulative probability $p(V) = P(|V| > |V_{th^{+/-}}|)$ for a voltage drop of V across the device. This is expressed as

$$p(|V|) = \int_{-\infty}^{|V|} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-V_{th^{+/-}})^2}{2\sigma^2}} dx = 1 - Q(\frac{|V| - |V_{th^{+/-}}|}{\sigma}) \quad (1)$$

$V_{th^{+/-}}$ are the mean threshold voltages with σ as the standard deviation. In this simulation, we have chosen program and erase threshold means as $V_{th^+} = 0.1V$ and $V_{th^-} = -0.1V$ respectively with $\sigma = 0.1V$. The LRS conductance of a bistable RRAM is of 1 unit and HRS conductance is assumed to be zero.

In this setup, both axonal attenuations α_i and dendritic attenuations β_j are set to pre-selected varying attenuation in the range of 0.8 to 1. These produce 16 positive and 16 negative voltage levels shown in Figure 9 (a). Due to this staggering of pulse voltages, each RRAM experiences one a distinct switching probability as a function of Δt as in Figure 9 (b). Figures 9 (c,d) show the STDP learning windows with normalized change in the conductance (Δw) of the compound synapse, with and without dendrites.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 September 2018 doi:10.20944/preprints201807.0362.v2

Peer-reviewed version available at J. Low Power Electron. Appl. 2018, 8, 34; doi:10.3390/jlpea8040034

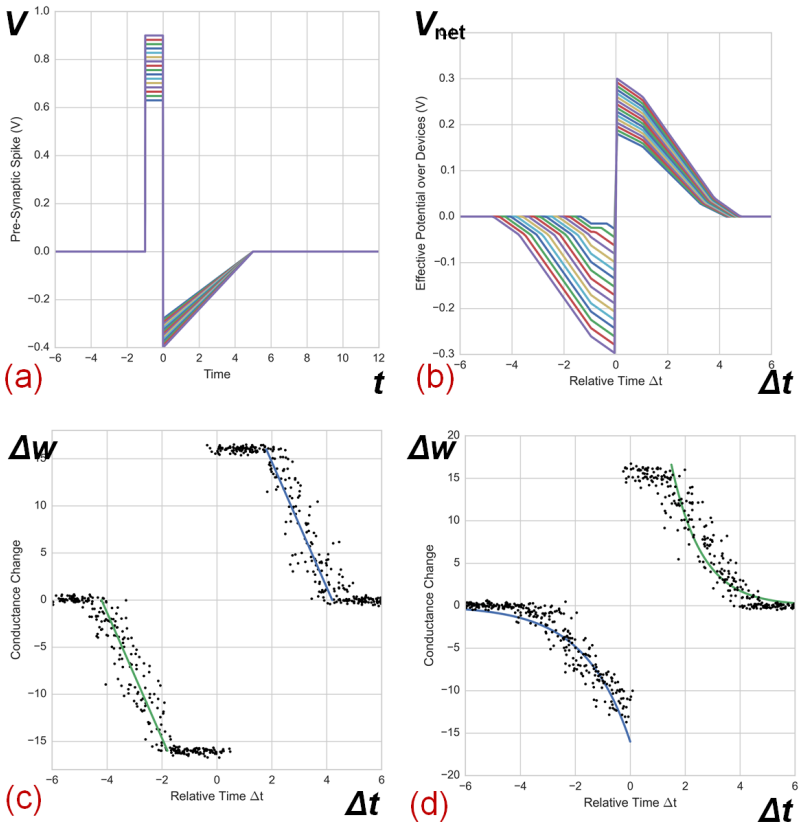

Figure 9. (a) Simulated spike waveforms with dendritic attenuations. (b) Effective potential difference V_{eff} across parallel devices versus Δt; 16 levels are created over program and erase thresholds V_{th+} and V_{th-}. Simulated STDP learning window without (c) dendrites and with (d) dendritic processing.

The plots in Figure 9 (c, d) demonstrate 16 levels of $\Delta w = \Delta G$, both in the LTP (positive) and LTD (negative) side of the STDP window. These 16 levels result in 4-bit resolution on average. Each dot in the plots represents the probability density of the particular Δw transition between -16 and 16. In Figure 9 (c), without dendrites, a linear curve fits the positive and negative sides of the resulting STDP window. And in Figure 9 (d), with dendritic processing, a double exponential curve is fitted to the simulated STDP window with <1-unit fitting error; the STDP window in (c) without dendrites has approximately 4-units error when fitted to the double exponential.

Moreover, the axonal and dendrite coefficients, α_i and β_j, and potentially their time delays, can be customized to implement a wide range of STDP learning windows. In additional to including nonlinear processing in future work, tuning of coefficients during training can allow mimicry of neuromodulation effects, where global error feedback signals can modulate local synaptic plasticity. Thus, in summary, combining dendritic processing schemes with stochastic RRAMs can allow high degree of freedom in implementing high-resolution STDP weights.

4.2. Modified CMOS Neuron with Dendritic Processing

An event-driven integrate-and-fire neuron circuit is adapted from the discussion in Section 3 and shown in Figure 10. Here, dendritic processing is realized by allowing parallel outputs with different gains/attenuations. The dendrites can be implemented using self-biased source follower (SF) based buffers with varying attenuations. The output impedance of the source follower buffers is designed to be smaller than the equivalent LRS resistance of the devices in parallel (R_{LRS}/M). Since the buffers external to the opamp in the CMOS neuron drive the resistive synapses, the power consumption of the opamp is considerably reduced. Consequently, single-stage opamp with $\approx 40dB$ gain and large

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 September 2018 doi:10.20944/preprints201807.0362.v2

Peer-reviewed version available at J. Low Power Electron. Appl. 2018, 8, 34; doi:10.3390/jlpea8040034

input swing is sufficient to realize the neuron. Furthermore, by splitting the buffers needed to drive the RRAM synapses for each dendrite allows larger synaptic fan-outs. The pre-synaptic buffers in the axonal path, needed for backpropagating the spikes, require some thought. During the integration phase, these buffers should allow the input current to be summed at the opamp's virtual ground and integrated in the membrane capacitance. Thus, the axonal buffers are bypassed when the neuron is in the integration phase as shown in Figure 10. In future, nonlinearity in the dendritic circuits can be explored for realizing higher resolution with bistable RRAM synapses, as observed in neurobiology experiments.

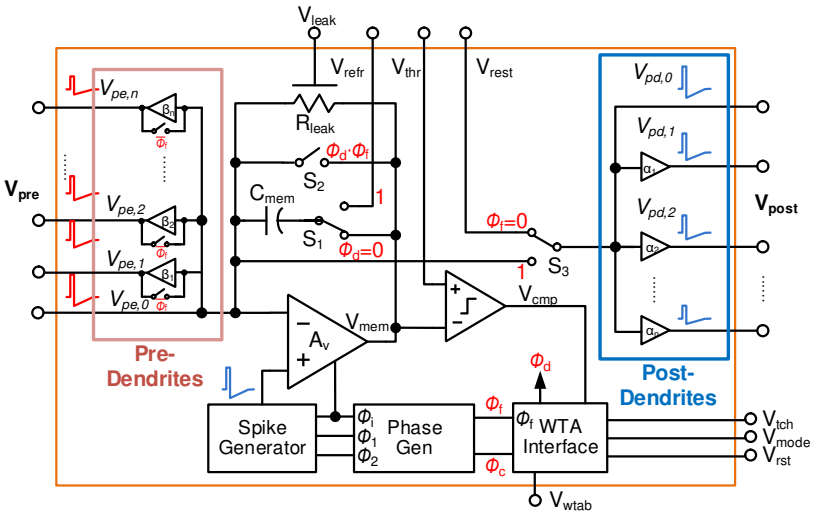

Figure 10. A simplified schematic of a spiking CMOS Neuron modified to accommodate pre-synaptic axonal and post-synaptic dendritic attenuations.

5. Energy-efficiency of Neuromorphic SoCs

The fundamental reason for investigating NVM or RRAM based NeuSoC architectures is to realize several order of magnitude improvement in energy-efficiency over the digital ASICs and GPUs, and neurosynaptic density when compared to contemporary neuromorphic chips that use digital SRAM synapses. As discussed earlier, resistive loading of CMOS neural circuits by the resistive synapses poses severe limitations on the achievable energy-efficiency of the NeuSoCs. In the proposed NeuSoC architecture, two factors primarily determine the energy-efficiency at the circuit level, namely the spike shape parameters (voltage and pulse width) and the range of the low-resistance state, R_{LRS}. The high-resistance state range R_{HRS} is typically order(s) of magnitude higher than R_{LRS} an thus can be ignored in the energy-efficiency calculations. The spike pulse shapes are as seen in the Fig. 3. The spike pulse-shape has an amplitude A^+ and pulse-width of τ^+ during the inference mode. Thus, the current input signal is $I_{syn} = \frac{A^+}{R_M}$ and the energy consumed in driving a synapse with resistance R_M, $\frac{R_{LRS}}{M} < R_M < \frac{R_{HRS}}{M}$, is

$$E_{spk} = \frac{A^{+2}\tau^+}{R_M} < \frac{A^{+2}\tau^+}{R_{LRS}} \tag{2}$$

In this calculation, compound synapses with dendritic processing and $M = 16$ RRAM devices in parallel are employed to achieve an equivalent analog synapse with 4-bit resolution, as discussed earlier in Section 4.1. Learning algorithmic considerations such as the input encoding, neuron sparsity (i.e. the percentage of synapses in LRS state), neuron spike-rate adaptation and homeostasis also determine the energy-efficiency of the overall NeuSoC. P_{neuron} is the neuron static power consumption and their sparsity factor $\eta_{sparsity}$. For a single instance of training or inference on an input pattern, the energy consumed in the spiking neural network is approximated as

$$E_{SNN} = \eta_{sp}\eta_{LRS}N_sE_{spk} + N_nP_n\tau^+ \quad (3)$$

where N_n is the total number of neurons, N_s is the total number of synaptic connections in the SNN, η_{LRS} is the fraction of synapses in the LRS-state. Furthermore, energy dissipated in the peripheral circuits outside the neurosynpatic array is ignored for the benefit of analytical simplicity.

For benchmarking the NeuSoC architecture performance, we employ the AlexNet deep convolutional neural network (CNN) that was the winner of the Imagenet Computer Vision Challenge [58]. The Alexnet neural network was trained on then state-of-the-art Nvidia P4 GPU, had 640k neurons with 61 million synapses and had a classification energy-efficiency of 170 images/second/Watt [59]. Here, we envisage an equivalent SNN that can achieve classification accuracy within 1% error as that of the deep neural network trained on a GPU. We assume that this is achievable by using transfer learning [44] in an SNN, and/or by employing spike-based equivalent of the backpropagation or similar algorithms [48]. Thus, the overall circuit architecture is essentially the same as that of the standard ANN, but implemented using mixed-signal neurons and RRAM synapses.

The numerical estimates are based on our previously reported spiking neural chip designed to drive RRAM synapses With an estimation based on the RRAM-compatible spiking neuron chip realized in [39], 4-bit compound memristive synapses [35,56,57], and R_{LRS} ranging from 0.1-10$M\Omega$, the energy consumption for processing (training or classification) of one image is shown in Table 1. By comparing with the contemporary GPU Nvidia P4 [59] (170 images/s/W), a memristive architecture with $R_{LRS} = 100k\Omega$ provides a meager 14$\times$ improvement in energy-efficiency. However, the energy consumption can be significantly reduced if the LRS resistance of the memristive devices can be increased to high-$M\Omega$ regime, leading to a potential 1000$\times$ range performance improvement; high LRS also helps reduce the power consumption in the opamp-based neuron circuits [39,60]. This analysis suggests that the energy-efficiency can be improved solely by increasing the LRS resistance of the RRAM devices.

Table 1. Energy estimation for a NeuSoC employing compound RRAM synapse with M=16 parallel devices.

		Low	**Medium**	**High**
Spike Width	τ^+		100ns	
Spike Amplitude	A^+		300mV	
LRS Resistance	R_{LRS}	100$k\Omega$	1$M\Omega$	10$M\Omega$
Single Spike Energy	E_{spk}	1.4pJ	140fJ	14fJ
Neuron Energy	E_N	1.56pJ	260fJ	43.3fJ
Neuron Sparsity	η_{sp}		0.6	
Fraction of RRAMs in LRS	η_{LRS}		0.5	
Single Event Energy	E_{SNN}	422.6μJ	42.33μJ	4.24μJ
Images/sec/watt		2.4k	23.6k	235k
Acceleration over GPU		$\times$14	$\times$139	$\times$1.38k

6. Towards Large Scale Neuromorphic SoCs

We have described the underlying device design and operation trade-offs for the emerging memory devices in NeuSoC applications. The write (Program/Erase) and read pulse voltages and temporal profile govern the fundamental tradeoffs between performance parameters such as the state retention, stochasticity, crossbar array size and impact of sneak-paths, device endurance, and energy consumption. The LRS resistance governs the energy-efficiency of the NeuSoC. However, the synapses resistance range trades-off with the available signal-to-noise ratio (SNR) during inference, as a higher HRS would result in the current being integrated to be of the same order as the thermal and flicker noise in the CMOS neuron. The synapse resistance range (or the HRS/LRS ratio),

synapse stochasticity, and the inference SNR ultimately determine the learning and classification performance of the deep learning architectures. For example, we may require higher endurance if the NeuSoC continually trains while in operation, or the NeuSoC is desired for continual use in real-time computing for several years. This may require applying lower stress to the devices which can result in higher stochasticity. The amount of stochasticity directly impacts the state retention (more state leakage or relaxation for higher stochasticity). Thus, it's imperative that the device optimization cannot be decoupled from the application-level circuit and system-level requirements.

Further, stochastity provides a viable approach for multibit synpase realization using realistic devices. In the near term, the crossbar circuit architecture will continually advance to realize ConvNets and implement the emerging learning algorithms where error feedback (such as in backprop) can be implemented using evolving mechanisms such as neuromodulated STDP, random backpropagation [38], or through explicit computation of gradients. Continuing and future work entails a closed-loop development paradigm where a device probing testbed is designed with certain application-oriented operating parameters in mind. Here, fabricated devices will be characterized for the spiking pulse profiles needed for accomplishing a system-level performance metrics. Then, these parameters will be plugged into a system-scale simulation (in Python) to predict the impact on the overall classification performance.

7. Conclusion

This work provides a review of the application of RRAM synapses to mixed-signal neuromorphic computing and challenges involved in their interface with CMOS neuron circuits. The interplay of devices, circuits and algorithm is important and their co-development is critical in optimizing the overall energy-efficiency of the NeuSoC architecture and bringing it closer to the biology-like efficiency. With continued progress, such neuromorphic architectures pave the path for computing beyond the limitations set by the Moore's scaling of CMOS transistors and the energy bottleneck of von Neumann computers. Moreover, such NeuSoCs open the possibility of realizing general purpose Artificial Intelligence in portable devices instead of always relying upon the energy-intensive Cloud infrastructure. In doing so, NeuSoCs provide newer avenue for memory technology development, where memory itself can be the next generation platform, integral to computing. In-memory computation occurring in these NeuSoC architectures will place the emphasis on dense integration of memory arrays with peripheral neural circuits, extending to 3D stacking and network on chip. Future work includes simulation and evaluation of device parameters for simultaneous development and fine-tuning of learning algorithms for targeted applications.

Acknowledgments: The authors gratefully acknowledge partial support through NSF award EECS-1454411 and Micron Foundation for the endowment. The authors also thank Prof. Maria Mitkova for discussions on CBRAMs and Prof. John Chiasson for technical discussions on spiking neural networks.

Author Contributions: For research articles with several authors, a short paragraph specifying their individual contributions must be provided. The following statements should be used "X.W. and V.S. conceived and designed the experiments; X.W. performed the experiments; X.W. and V.S. analyzed the data; I.S. contributed analysis on bio-plausibility of methods; K.Z. helped with chip design and test. V.S. coordinated writing the paper.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ASIC: Application Specific Integrated Circuit
CMOS: Complementary metal oxide semiconductor
NVRAM: Non-volatile random access memory
RRAM: Resistive random access memory

SNN: Spiking neural networks
STDP: Spike-timing dependent plasticity
NeuSoC: Neuromorphic System-on-a-Chip

Bibliography

1. Williams, R.; DeBenedictis, E.P. OSTP Nanotechnology-Inspired Grand Challenge: Sensible Machines. *IEEE whitepaper, http://rebootingcomputing. ieee. org/archived-articles-and-videos/general/sensible-machine* **2015**.
2. Hinton, G.E.; Salakhutdinov, R.R. Reducing the dimensionality of data with neural networks. *Science* **2006**, *313*, 504–507.
3. Krzanich, B. Intel Pioneers New Technologies to Advance Artificial Intelligence.
4. Bi, G.q.; Poo, M.m. Synaptic modification by correlated activity: Hebb's postulate revisited. *Annual review of neuroscience* **2001**, *24*, 139–166.
5. Dan, Y.; Poo, M.m. Spike timing-dependent plasticity of neural circuits. *Neuron* **2004**, *44*, 23–30.
6. Masquelier, T.; Thorpe, S.J. Unsupervised Learning Of Visual Features Through Spike Timing Dependent Plasticity. *PLoS computational biology* **2007**, *3*, e31.
7. Nessler, B.; Pfeiffer, M.; Buesing, L.; Maass, W. Bayesian computation emerges in generic cortical microcircuits through spike-timing-dependent plasticity. *PLoS computational biology* **2013**, *9*, e1003037.
8. Merolla, P.A.; Arthur, J.V.; Alvarez-Icaza, R.; Cassidy, A.S.; Sawada, J.; Akopyan, F.; Jackson, B.L.; Imam, N.; Guo, C.; Nakamura, Y.; Brezzo, B.; Vo, I.; Esser, S.K.; Appuswamy, R.; Taba, B.; Amir, A.; Flickner, M.D.; Rish, W.P.; Manohar, R.; Modha, D.S. A million spiking-neuron integrated circuit with a scalable communication network and interface. *Science Magazine* **2014**, *345*, 668–673.
9. Painkras, E.; Plana, L.; Garside, J.; Temple, S.; Davidson, S.; Pepper, J.; Clark, D.; Patterson, C.; Furber, S. Spinnaker: a multi-core system-on-chip for massively-parallel neural net simulation. Custom Integrated Circuits Conference (CICC), 2012 IEEE. IEEE, 2012, pp. 1–4.
10. Davies, M.; Srinivasa, N.; Lin, T.H.; Chinya, G.; Cao, Y.; Choday, S.H.; Dimou, G.; Joshi, P.; Imam, N.; Jain, S.; others. Loihi: A neuromorphic manycore processor with on-chip learning. *IEEE Micro* **2018**, *38*, 82–99.
11. Boahen, K. Neurogrid: Emulating A Million Neurons In The Cortex. International Conference of the IEEE Engineering in Medicine and Biology Society, 2006.
12. Indiveri, G.; Chicca, E.; Douglas, R. A VLSI array of low-power spiking neurons and bistable synapses with spike-timing dependent plasticity. *IEEE Transactions On Neural Networks* **2006**, *17*, 211–21.
13. Neftci, E.; Das, S.; Pedroni, B.; Kreutz-Delgado, K.; Cauwenberghs, G. Event-driven contrastive divergence for spiking neuromorphic systems. *Frontiers in neuroscience* **2013**, *7*.
14. Brink, S.; Nease, S.; Hasler, P. Computing with networks of spiking neurons on a biophysically motivated floating-gate based neuromorphic integrated circuit. *Neural Networks* **2013**.
15. Lu, J.; Young, S.; Arel, I.; Holleman, J. A 1 TOPS/W analog deep machine-learning engine with floating-gate storage in 0.13 μm CMOS. *IEEE Journal of Solid-State Circuits* **2015**, *50*, 270–281.
16. Burr, G.W.; Shelby, R.M.; Sebastian, A.; Kim, S.; Kim, S.; Sidler, S.; Virwani, K.; Ishii, M.; Narayanan, P.; Fumarola, A.; Sanches, L.L.; Boybat, I.; Le Gallo, M.; Moon, K.; Woo, J.; Hwang, H.; Leblebici, Y. Neuromorphic computing using non-volatile memory. *Advances in Physics: X* **2017**, *2*, 89–124.
17. Kim, S.; Ishii, M.; Lewis, S.; Perri, T.; BrightSky, M.; Kim, W.; Jordan, R.; Burr, G.; Sosa, N.; Ray, A.; others. NVM neuromorphic core with 64k-cell (256-by-256) phase change memory synaptic array with on-chip neuron circuits for continuous in-situ learning. IEEE International Electron Devices Meeting (IEDM). IEEE, 2015, pp. 17–1.
18. Burr, G.W.; Shelby, R.M.; Sidler, S.; Di Nolfo, C.; Jang, J.; Boybat, I.; Shenoy, R.S.; Narayanan, P.; Virwani, K.; Giacometti, E.U.; others. Experimental demonstration and tolerancing of a large-scale neural network (165 000 synapses) using phase-change memory as the synaptic weight element. *IEEE Transactions on Electron Devices* **2015**, *62*, 3498–3507.
19. Jo, S.H.; Chang, T.; Ebong, I.; Bhadviya, B.B.; Mazumder, P.; Lu, W. Nanoscale memristor device as synapse in neuromorphic systems. *Nano letters* **2010**, *10*, 1297–1301.
20. Li, Y.; Zhong, Y.; Xu, L.; Zhang, J.; Xu, X.; Sun, H.; Miao, X. Ultrafast Synaptic Events In A Chalcogenide Memristor. *Scientific Reports* **2013**, *3*.

21. Yang, J.J.; Strukov, D.B.; Stewart, D.R. Memristive devices for computing. *Nature nanotechnology* **2013**, *8*, 13–24.
22. Chang, T.; Yang, Y.; Lu, W. Building neuromorphic circuits with memristive devices. *IEEE Circuits and Systems Magazine* **2013**, *13*, 56–73.
23. Yu, S.; Kuzum, D.; Wong, H.S.P. Design considerations of synaptic device for neuromorphic computing. IEEE International Symposium on Circuits and Systems (ISCAS). IEEE, 2014, pp. 1062–1065.
24. Indiveri, G.; Legenstein, R.; Deligeorgis, G.; Prodromakis, T. Integration Of Nanoscale Memristor Synapses In Neuromorphic Computing Architectures. *Nanotechnology* **2013**, *24*, 384010.
25. Wu, X.; Saxena, V.; Zhu, K. Homogeneous Spiking Neuromorphic System for Real-World Pattern Recognition. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems (JETCAS)* **2015**, *5*, 254 – 266.
26. Saxena, V. Memory Controlled Circuit System and Apparatus, 2015. US Patent App. 14/538,600.
27. Saxena, V. A Compact CMOS Memristor Emulator Circuit and its Applications. IEEE International Midwest Symposium on Circuits and Systems (MWSCAS), 2018, pp. 1–5.
28. Saxena, V.; Wu, X.; Zhu, K. Energy-Efficient CMOS Memristive Synapses for Mixed-Signal Neuromorphic System-on-a-Chip. 2018 IEEE International Symposium on Circuits and Systems (ISCAS), 2018, pp. 1–5.
29. Govoreanu, B.; Kar, G.; Chen, Y.; Paraschiv, V.; Kubicek, S.; Fantini, A.; Radu, I.; Goux, L.; Clima, S.; Degraeve, R.; others. $10\times$ 10nm 2 Hf/HfO x crossbar resistive RAM with excellent performance, reliability and low-energy operation. IEEE International Electron Devices Meeting (IEDM). IEEE, 2011, pp. 31–6.
30. Chen, Y.Y.; Degraeve, R.; Clima, S.; Govoreanu, B.; Goux, L.; Fantini, A.; Kar, G.S.; Pourtois, G.; Groeseneken, G.; Wouters, D.J.; others. Understanding of the endurance failure in scaled HfO 2-based 1T1R RRAM through vacancy mobility degradation. IEEE International Electron Devices Meeting (IEDM), 2012, pp. 20–3.
31. Kozicki, M.N.; Mitkova, M.; Valov, I., Electrochemical Metallization Memories. In *Resistive Switching*; Wiley-Blackwell, 2016; pp. 483–514.
32. Fong, X.; Kim, Y.; Venkatesan, R.; Choday, S.H.; Raghunathan, A.; Roy, K. Spin-transfer torque memories: Devices, circuits, and systems. *Proceedings of the IEEE* **2016**, *104*, 1449–1488.
33. Micron. 3D XPoint^{TM} Technology: Breakthrough Nonvolatile Memory Technology.
34. Strukov, D.B.; Snider, G.S.; Stewart, D.R.; Williams, R.S. The Missing Memristor Found. *Nature* **2008**, *453*, 80.
35. Saxena, V.; Wu, X.; Srivastava, I.; Zhu, K. Towards spiking neuromorphic system-on-a-chip with bio-plausible synapses using emerging devices. Proceedings of the 4th ACM International Conference on Nanoscale Computing and Communication. ACM, 2017, p. 18.
36. Kuzum, D.; Jeyasingh, R.G.; Lee, B.; Wong, H.S.P. Nanoelectronic programmable synapses based on phase change materials for brain-inspired computing. *Nano letters* **2011**, *12*, 2179–2186.
37. Seo, K.; Kim, I.; Jung, S.; Jo, M.; Park, S.; Park, J.; Shin, J.; Biju, K.P.; Kong, J.; Lee, K.; Lee, B.; Hwang, H. Analog memory and spike-timing-dependent plasticity characteristics of a nanoscale titanium oxide bilayer resistive switching device. *Nanotechnology* **2011**, *22*, 254023.
38. Koch, C. Computation and the single neuron. *Nature* **1997**, *385*, 207.
39. Wu, X.; Saxena, V.; Zhu, K.; Balagopal, S. A CMOS Spiking Neuron for Brain-Inspired Neural Networks With Resistive Synapses and In Situ Learning. *IEEE Transactions on Circuits and Systems II: Express Briefs* **2015**, *62*, 1088–1092.
40. Wu, X.; Saxena, V.; Zhu, K. A CMOS Spiking Neuron For Dense Memristor-synapse Connectivity For Brain-inspired Computing. International Joint Conference on Neural Networks (IJCNN), 2015, pp. 1–6.
41. Latif, M.R. Nano-Ionic Redox Resistive RAM–Device Performance Enhancement Through Materials Engineering, Characterization and Electrical Testing. PhD thesis, 2014.
42. Masquelier, T.; Guyonneau, R.; Thorpe, S.J. Spike Timing Dependent Plasticity Finds The Start Of Repeating Patterns In Continuous Spike Trains. *PloS One* **2008**, *3*, e1377.
43. Diehl, P.U.; Cook, M. Unsupervised Learning Of Digit Recognition Using Spike-timing-dependent Plasticity. *Frontiers in Computational Neuroscience* **2015**, *9*.

44. Diehl, P.U.; Neil, D.; Binas, J.; Cook, M.; Liu, S.C.; Pfeiffer, M. Fast-classifying, high-accuracy spiking deep networks through weight and threshold balancing. International Joint Conference onNeural Networks (IJCNN), 2015, pp. 1–8.
45. Lee, J.H.; Delbruck, T.; Pfeiffer, M. Training deep spiking neural networks using backpropagation. *Frontiers in Neuroscience* **2016**, *10*.
46. Kheradpisheh, S.R.; Ganjtabesh, M.; Thorpe, S.J.; Masquelier, T. STDP-based spiking deep convolutional neural networks for object recognition. *Neural Networks* **2018**, *99*, 56–67.
47. Tavanaei, A.; Maida, A.S. Bio-Inspired Spiking Convolutional Neural Network using Layer-wise Sparse Coding and STDP Learning. *arXiv preprint arXiv:1611.03000* **2016**.
48. Neftci, E.O.; Augustine, C.; Paul, S.; Detorakis, G. Event-driven random back-propagation: Enabling neuromorphic deep learning machines. *Frontiers in neuroscience* **2017**, *11*, 324.
49. He, W.; Sun, H.; Zhou, Y.; Lu, K.; Xue, K.; Miao, X. Customized binary and multi-level HfO 2- x-based memristors tuned by oxidation conditions. *Scientific Reports* **2017**, *7*, 10070.
50. Beckmann, K.; Holt, J.; Manem, H.; Van Nostrand, J.; Cady, N.C. Nanoscale Hafnium Oxide RRAM Devices Exhibit Pulse Dependent Behavior and Multi-level Resistance Capability. *MRS Advances* **2016**, *1*, 3355–3360.
51. Kandel, E.R.; Schwartz, J.H.; Jessell, T.M.; Siegelbaum, S.A.; Hudspeth, A.J. *Principles of Neural Science*; Vol. 4, McGraw-Hill New York, 2000.
52. Pedrosa, V.; Clopath, C. The Role of Neuromodulators in Cortical Plasticity. A Computational Perspective. *Frontiers in synaptic neuroscience* **2016**, *8*.
53. Poirazi, P.; Mel, B.W. Impact of active dendrites and structural plasticity on the memory capacity of neural tissue. *Neuron* **2001**, *29*, 779–796.
54. Rueckauer, B.; Lungu, I.A.; Hu, Y.; Pfeiffer, M.; Liu, S.C. Conversion of continuous-valued deep networks to efficient event-driven networks for image classification. *Frontiers in neuroscience* **2017**, *11*, 682.
55. Pfeil, T.; Potjans, T.C.; Schrader, S.; Potjans, W.; Schemmel, J.; Diesmann, M.; Meier, K. Is a 4-bit synaptic weight resolution enough?-constraints on enabling spike-timing dependent plasticity in neuromorphic hardware. *arXiv preprint arXiv:1201.6255* **2012**.
56. Bill, J.; Legenstein, R. A compound memristive synapse model for statistical learning through STDP in spiking neural networks. *Frontiers in neuroscience* **2014**, *8*.
57. Wu, X.; Saxena, V. Enabling Bio-Plausible Multi-level STDP using CMOS Neurons with Dendrites and Bistable RRAMs. International Joint Conference on Neural Networks (IJCNN), Alaska, USA, 2017.
58. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet Classification With Deep Convolutional Neural Networks. Advances In Neural Information Processing Systems, 2012, pp. 1097–1105.
59. Nvidia. New Pascal GPUs Accelerate Inference in the Data Center, 2016.
60. Saxena, V.; Baker, R.J. Indirect Compensation Techniques For Three-stage Cmos Op-amps. IEEE International Midwest Symposium on Circuits and Systems (MWSCAS), 2009, pp. 9–12.