 doi:10.20944/preprints201807.0362.v1

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 a 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 CMOS mixed-signal integrated circuits and nanoscale emerging memory devices can enable a new generation of Neuromorphic computers that will alleviate 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 reduction in energy consumption. Practical demonstration of such architectures has been impeded as the performance of these emerging devices falls short of the expected behavior from the idealized analog synapses, or weights, and new 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: CMOS Neurons, Cognitive Computing, Deep Learning, Neuromorphic System-on-a-Chip (NeuSoC), NVRAM, RRAM, Spiking Neural Networks (SNNs).

1. Introduction

In 2015, the U.S. Office of Science and Technology (OSTP) announced a nanotechnology-inspired grand challenge that urged 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 take long computing cluster days

 doi:10.20944/preprints201807.0362.v1

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

Progress in neuromorphic hardware platforms has led to the realization asynchronous event-driven, as opposed to clock-driven, computers that communicate information across the chips using voltage spikes. 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.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 July 2018 doi:10.20944/preprints201807.0362.v1

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 Devices

Other solutions include using floating gate devices, or Flash memory, for implementing non-volatile synaptic weights [14,15]. However, 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 inference-only applications or scenarios where learning concludes within the endurance limit of the devices.

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 ($Terabits/cm^2$) [16]. These devices have demonstrated biologically plausible STDP plasticity behavior in several experiments [16,17], 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) [18–20]. Following this trend, hybrid CMOS-RRAM analog very-large-scale integrated (VLSI) circuits have been proposed [21,22] 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 [23,24], and resulting dynamic synapse circuits [25] 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.

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) [26,27], phase-change

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 July 2018 doi:10.20944/preprints201807.0362.v1

memory, conductive-bridge RAM (CBRAM) [28], STTRAM [29], and 3D crosspoint memory [30]. These are also referred as memristors or memristive devices in literature [16,31].

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 [22,32]. 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.

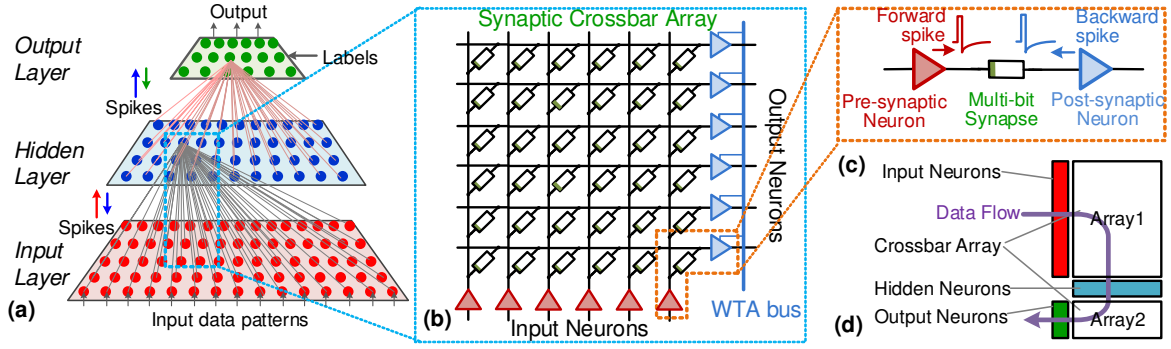

Figure 1. Neuromorphic SoC architecture: (a) A fully-connected spiking neural network showing input, hidden and output layers comprised of spiking neurons. Here, synaptic connections shown for one neuron in the hidden and output layers; (b) A slice of the neural network architecture implemented using 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 [20,33,34], 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 [35]." 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 (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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 July 2018 doi:10.20944/preprints201807.0362.v1

to enable on-chip learning. This not only saves energy but also avoids undesirable changes to the synaptic weights.

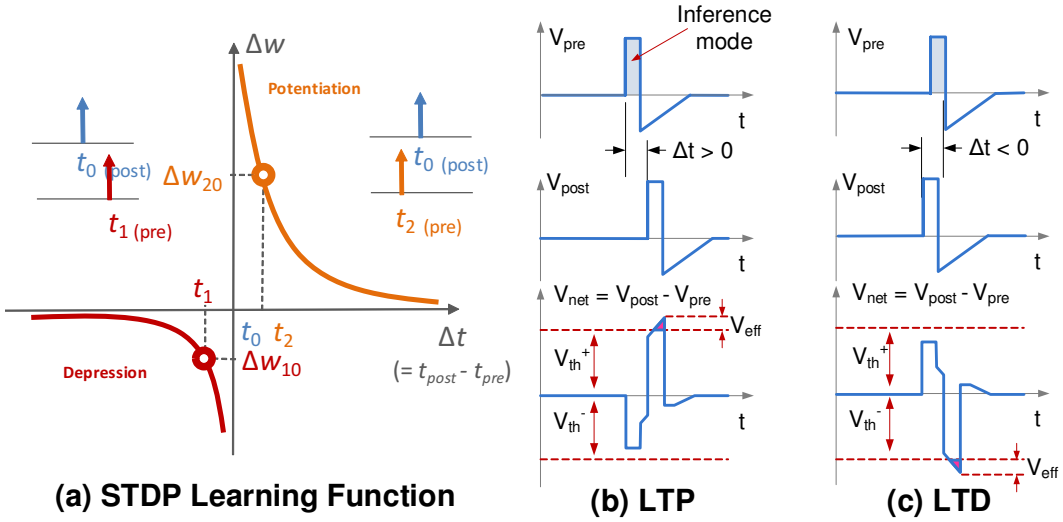

Figure 2. A two-layer spiking neural network crossbar RRAM synapse array and peripheral CMOS neurons with WTA bus. During the inference mode only the pre-spikes with the positive rectangular pulse are used; the post-spikes and the negative tails are activated during the training mode.

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 [36,37]. 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 [38]. 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 [22,38].

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: 19 July 2018 doi:10.20944/preprints201807.0362.v1

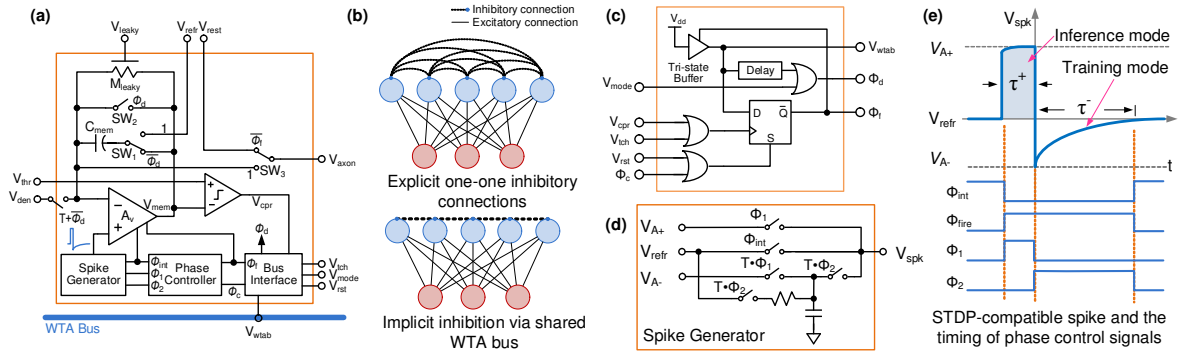

Figure 3. (a) Schematic of the integrate-and-fire Neuron for neural learning. (b) Competitive learning uses explicit one-on-one inhibitory connections, whereas the same function can be implemented with implicit inhibition on a shared WTA bus. (c) The proposed asynchronous WTA bus interface circuit. (d) 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) [22].

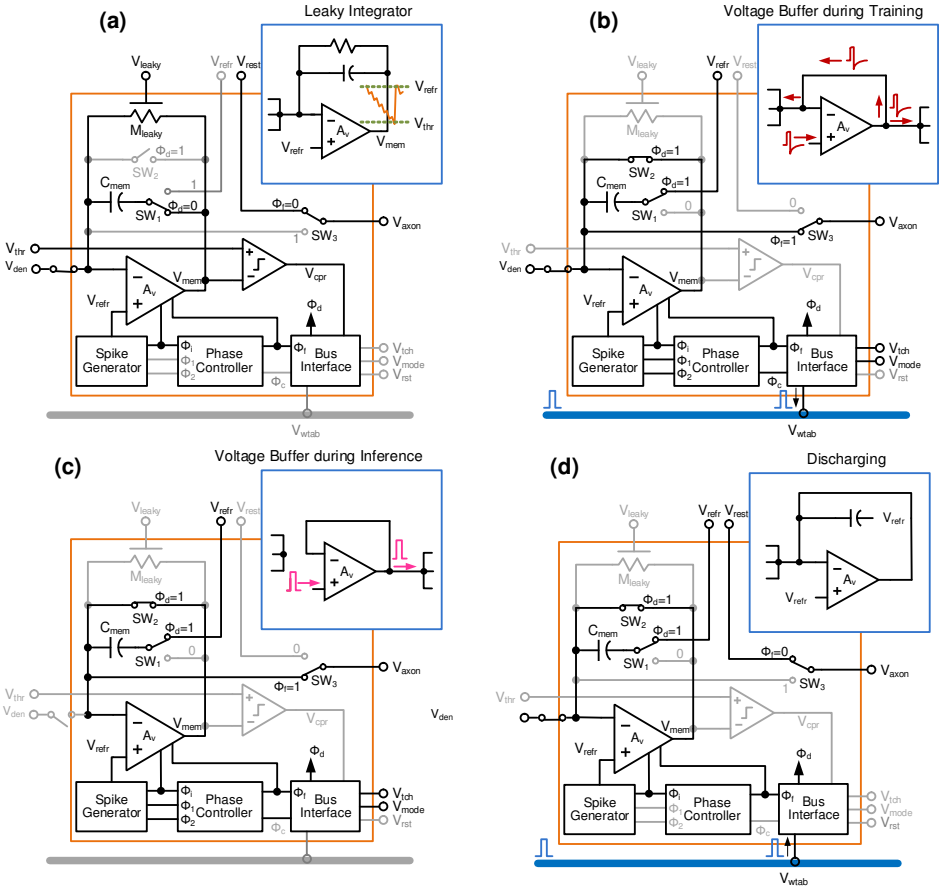

Figure 4. Event-driven operation of the proposed leaky integrate-and-fire neuron during training and inference.

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: 19 July 2018 doi:10.20944/preprints201807.0362.v1

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 [22]. A chip was designed using an earlier version of this neuron where associative learning (Pavlov's dog experiment) was demonstrated [38] using conductive bridge type resistive synaptic devices [39], as shown in Figure 5.

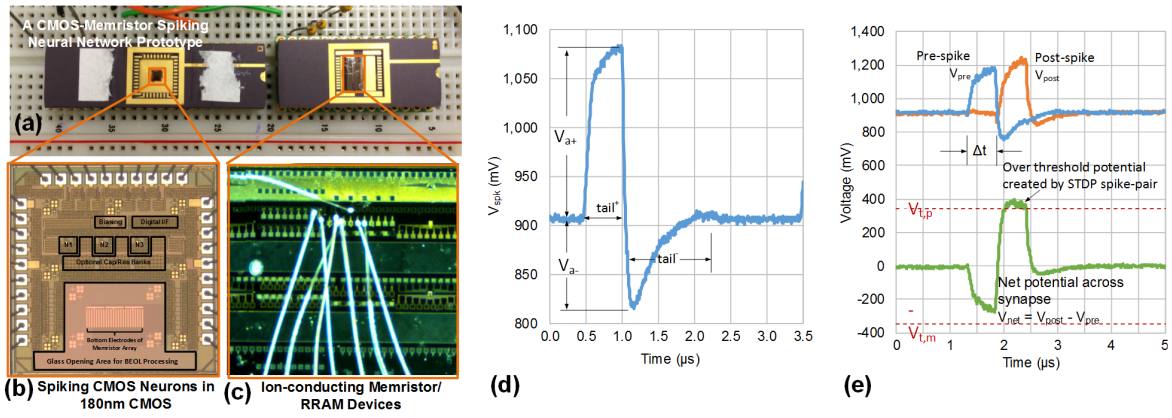

Figure 5. RRAM-compatible CMOS Neuron: (a) A CMOS-RRAM experimental prototype with (b) 180-nm CMOS spiking neuron chips with digital reconfigurability, and (c) CBRAM devices. (d) Measured spike output for one of the settings, (e) Pre- and post-spike voltage difference applied across a synapse [38].

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 [7,40]. 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 [22,37]. 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 [22]. 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%.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 July 2018 doi:10.20944/preprints201807.0362.v1

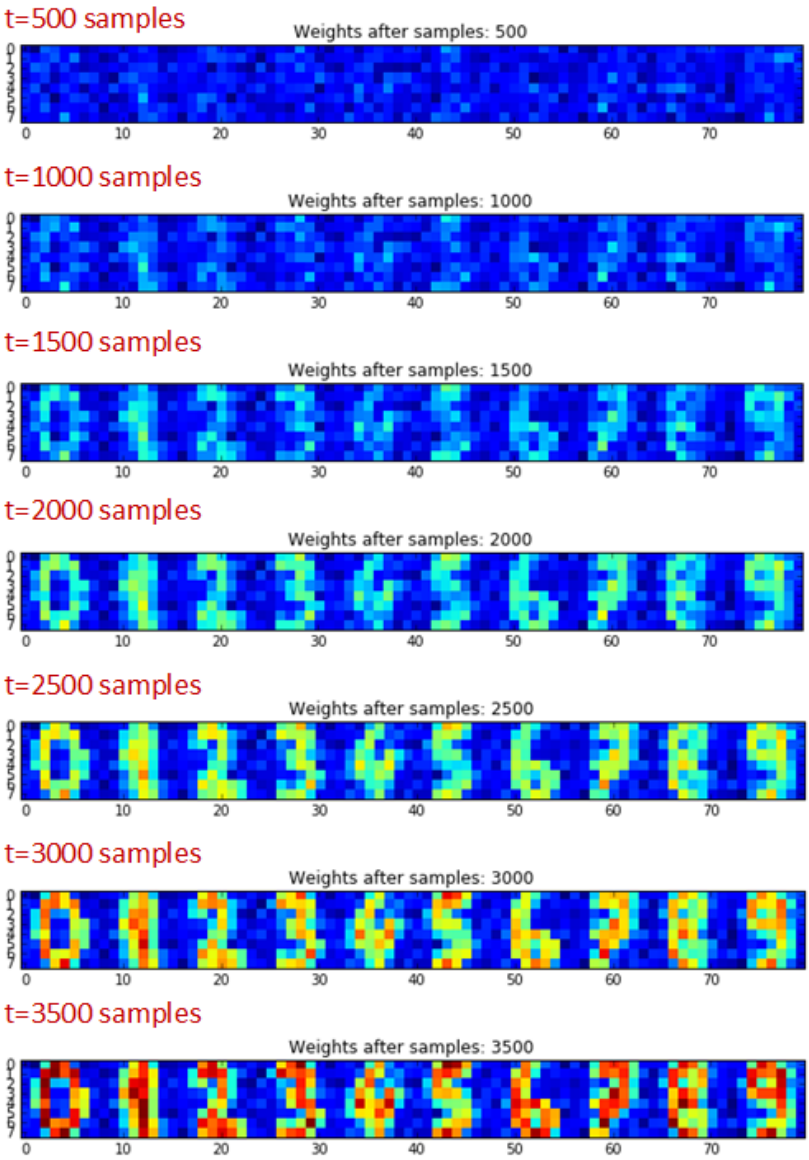

Figure 6. Evolution of synaptic weights (normalized to the color scale) in the SNN for 8 × 8 handwritten character classification.

Higher classification accuracy can be potentially achieved by increasing the number of competing neurons [41] 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 [42], followed by a spiking backprop that used membrane potential as a differentiable function [43]. In parallel, semi-supervised deep spike-based convolutional networks (ConvNets) for image pattern classification using GPUs have claimed >98% classification accuracy [44,45]. Moreover, there is a growing interest in developing backprop for deep SNNs with some success [46]. 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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 July 2018 doi:10.20944/preprints201807.0362.v1

algorithms must be re-casted based upon the behavior of the synaptic devices such as in [47], where the STDP was modified to accommodate abrupt reset (depression) in PCM-based synapses.

3.5. Challenges with Emerging Devices as Synapses

Current memristive or RRAM devices exhibit several practical limitations in building neuromorphic systems:

(1) Variability: Device characteristics such as the switching threshold voltages are variable, or stochastic, for each device and across several devices and depend upon the initial 'forming' step where the filament is formed in the RRAM. The Set (program) threshold required for filament formation varies with the compliance current (I_{CC}) which also determines the characteristics of the low-resistance state (LRS). For a small $I_{CC} = 50nA$, a weak filament is formed that exhibits analog-like behavior with a large variance in LRS resistance. However, these states may relax within minutes to several hours. For a larger $I_{CC} = 5\mu A$, a thick filament is formed that exhibits binary behavior with a small variance in LRS. It must be noted that the compliance current cannot be easily set in a crossbar array configuration due to large circuit overheads.

(2) Low Resistance: A large resistance is desirable in the 'On' or LRS state to reduce the power consumption in the neurons that drive these synapses. Ideally $10 - 100M\Omega$ LRS are desired as a trade-off between power and circuit noise; device stoichiometry and material selection should consider this constraint.

(3) Resolution and Retention: It has been a challenge to realize stable weights for more than single-bit resolution in RRAMs due to the relaxation of the filament forming mechanism. HfO_x and TaO_x based devices have exhibited $\approx$ 9 states and their stability in actual circuit is currently being investigated [48]. It is safe to assume that several of these devices can exhibit bistable behavior when used in a crossbar configuration without setting compliance current.

(4) Unipolarity: Several RRAM devices tend to be unipolar as only a small voltage is needed to break or dissolve the filament and send the device to a high resistance state (HRS). This is not compatible with the STDP scheme seen earlier in Figure 3.

(5) Endurance: Synaptic device endurance will determine the online or continuous learning ability in the neuromorphic chip. For example, floating-gate devices are best suited for inference due to $\approx 10^5$ cycles write endurance, while PCM devices can only be allowed to train, or learn, intermittently ($< 10^8$ write cycles).

4. Bio-inspiration for Higher-resolution Synapses

As the field of synaptic neuroscience develops, more insight is available in the underlying nature of synaptic plasticity and the underlying 'algorithm' in the brain. The synapses are formed where the axon terminals interconnect with the dendrites of the other neurons as shown in Figure 7. Each axon terminal contains thousands of membrane-bound vesicles, each containing thousands of neurotransmitter molecules. Neurotransmitters are biochemical molecules that relay and process signals between the neurons. When stimulated by an electrical spike at the pre-synaptic axon terminal, neurotransmitters of various types are released, and they are ejected from the cell membrane into the synaptic cleft between the neurons. These chemicals then bind to chemical receptors in the dendrites of the post-synaptic neuron. Consequently, they cause opening up of special gates which allow a flood of charged particles (Ca^{2+}, Na^{+}, K^{+} and Cl^{-}). This affects the potential charge of the post-neuron, which then creates an electrical current which is integrated in the receiving neuron. The whole process takes <2ms [49].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 July 2018 doi:10.20944/preprints201807.0362.v1

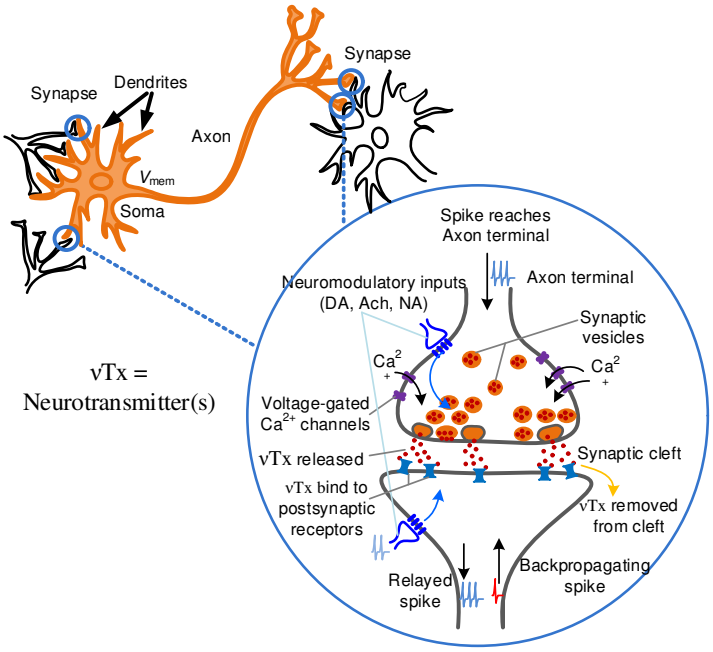

Figure 7. Synapse formed as the junction of axonal terminal and dendrites. Some of the signaling mechanisms are shown.

The backpropagating action potential (spike) influences the receptors by causing large Ca^{2+} transients. These trigger spike-timing based long-term potentiation (LTP) in the synapse. The long-term depression (LTD) mechanism is still not well understood. Neuromodulators such as dopamine (DA) and acetylcholine (Ach), etc., are known to impact plasticity. One principle neuromodulatory effect is to gate plasticity by modifying the spike-timing-dependent plasticity (STDP) learning window [50]. Other neuromodulation effects include regulation of neuronal activity and thus the learning rate. Understanding of neuromodulation will help uncover the presence of backprop-like or other unsupervised learning in the brain. Dendrites also play a role in neural signal processing through signal attenuation and potentially modification of STDP. Further, the neuromodulators release is a quantized and stochastic mechanism, and thus plasticity can also be surmised to be discrete and stochastic in nature; however, it has been difficult to verify the case or the contrary in experiments. Even if the synaptic weights were bistable, their averaging with dendritic filtering will provides semblance of analog and continuous STDP in neurobiology experiments.

As the computational neuroscience community improves its understanding of the underlying spike-domain signal processing occurring in biological synapses, their emulation in circuits will continue to be refined. Since the goal is to reduce the architecture to realize computing while discarding non-essential biological housekeeping and repair activities, we may discard the contributions of glia and astrocytes, which make up for other 50% of mass than the neurons, until their role in learning becomes clear.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 July 2018 doi:10.20944/preprints201807.0362.v1

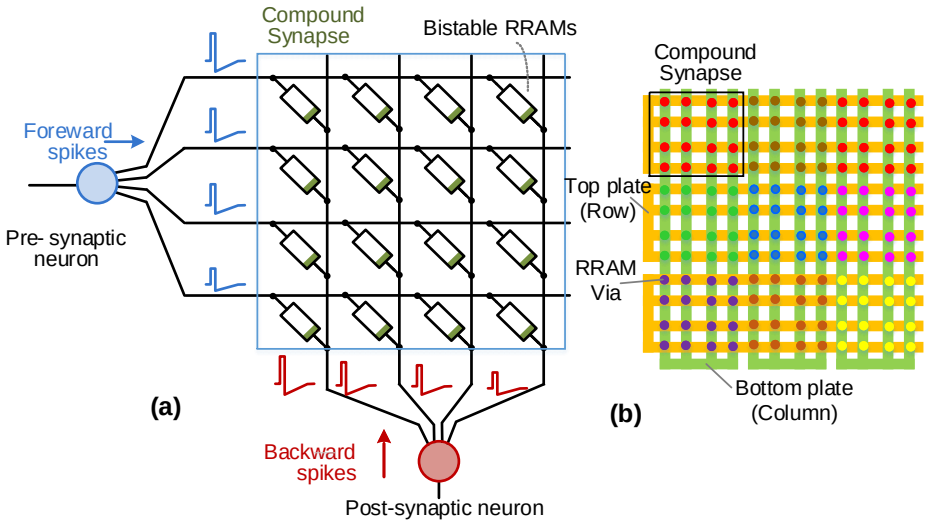

Figure 8. (a) A compound synapse with 16 bistable RRAMs in parallel with 4×4 dendritic configuration, (b) a possible layout configuration.

4.1. Compound Synapse with Dendritic Processing

We discussed the limitations of RRAM devices in realizing continuous analog STDP and high-resolution weights. Spiking neural network studies have established that $\geq$ 4 bits of synaptic resolution is required for meaningful tasks [51]. We can assume a worst-case scenario of bistable synapses. To circumvent these limitations, the authors introduced a new concept where multiple bistable synapses are combined leveraging their stochastic switching, with dendritic processing, to realize high-resolution plasticity [52]. Here, the concept is extended to pre- as well as post synaptic dendrites. This is illustrated in Figure 8 where several bistable RRAM devices are combined in the pre- and/or post-synaptic path to realize a compound synapse with multi-bit resolution with long-term storage. This is enabled by the fact that the devices are operated in stochastic switching regime and each device switches differently with respect to the time and/or voltage difference between the pre- and post-synaptic spikes (Δt).

The proposed concept is bio-plausible in the sense that the STDP updates are stochastic analogous to the probabilistic neurotransmitter signaling. The stochasticity in the forward spike transmission is inherently present due to the thermal noise and spike jitter in the neuron circuits. Further, averaging combination of several binary synapses with dendritic attenuation and delays, implements stochastic weight update that appears to be a continuous analog behavior. Here, LRS variation of RRAMs was modeled as a normal distributed random variable with a normalized standard deviation of 0.1. With dendritic processing applied to the pre- and post-synaptic spikes, the attenuating factors α_i and β_j of the dendritic attenuators were set to values non-linearly spanning from 0.8 to 1, and produced 16 positive and 16 negative levels. Simulation results in Figure 9 (c,d) show the STDP learning windows with normalized conductance changes with and without dendrites (a setup similar to [52] is used in these simulations). These plots clearly demonstrate an equivalent of 4-bits weight with 16 levels both on the positive as well as negative sides of the STDP window. Here, each dot represents the probability density of the state; a double exponential curve fits well to the simulated results (<1-unit fitting error) whereas the STDP curve without dendrites is better fitted to a linear curve (4-units fitting error to exponential). Further, individual dendrite coefficients, and potentially delay, can be tuned to realize a desirable STDP learning window shape. This opens the possibility of mimicking emerging mechanisms of cortical synapses, where dendrites and/or neuromodulators modulate local synaptic plasticity. These compound synapse variants pave the path towards robust embedded spiking deep learning architectures, even with low-resolution synaptic devices with high variability.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 July 2018 doi:10.20944/preprints201807.0362.v1

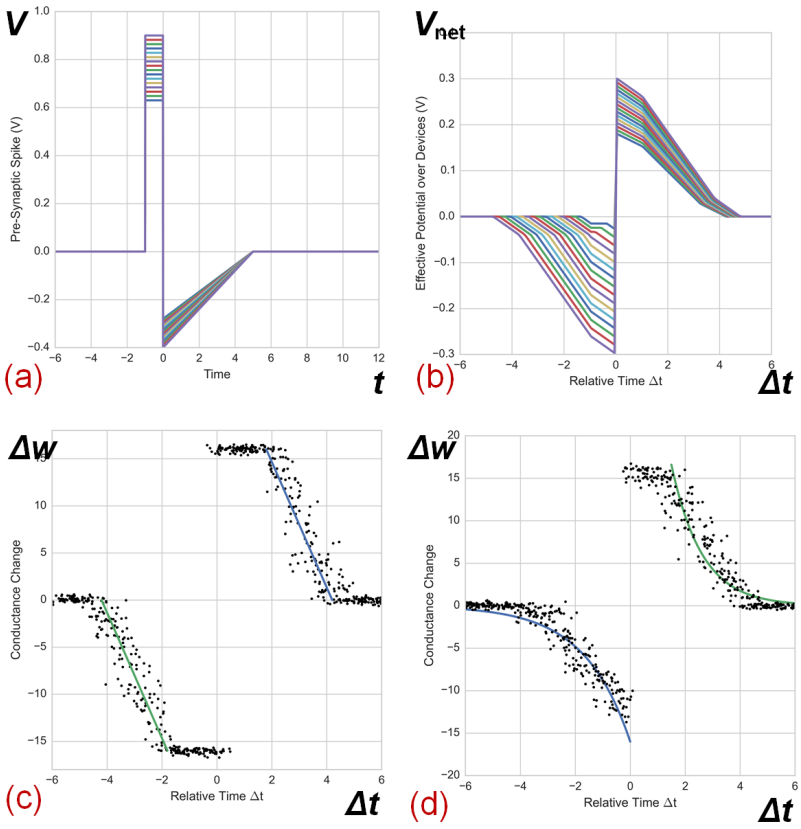

Figure 9. (a) Spike waveforms with dendritic attenuations. (b) Effective potential V_{eff} over parallel devices versus Δt; 16 levels are created over V_{th+} and V_{th-}. STDP learning window without (c) dendrites and with (d) dendritic processing.

4.2. Modified CMOS Neuron with Dendritic Processing

An event-driven integrate-and-fire neuron circuit is adapted from the discussion in Section 3. Here dendritic processing is included by allowing parallel outputs with different gain/attenuation. The dendrites are implemented using self-biased source follower (SF) based buffers with varying attenuation. The output impedance of the buffers is designed to be much smaller than the LRS resistance of the devices in parallel. Since buffers drive the resistive load, the power consumption of the opamp is considerably reduced in the firing phase; a single stage opamp with $\approx 40dB$ gain and large input swing is sufficient. Furthermore, splitting the buffers needed to drive the RRAM synapses, allows accommodation of the higher synaptic fanout necessitated by dendritic processing. The pre-synaptic dendrites are interesting; they should allow the input current to be summed at the opamp's virtual ground and integrated in the neuron. Thus, the buffers are bypassed when the neuron is in integrating phase. In order to accommodate unipolar synapses, such as the CBRAMs, a parallel diode structure (forward and reversed-based diodes realized using n-channel MOSFETs in parallel) can be employed at the neuronal dendrite output to simulate sufficiently large negative threshold for the synapses. In future, nonlinearity in the dendritic circuits can be explored for realizing higher resolution with bistable RRAM synapses, as observed in neurobiology experiments.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 July 2018 doi:10.20944/preprints201807.0362.v1

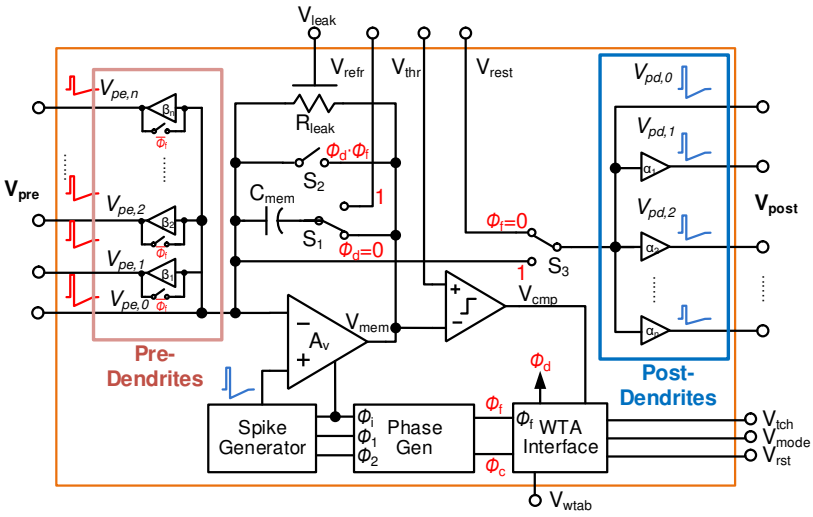

Figure 10. A simplified schematic of a spiking CMOS Neuron modified to accommodate pre- and post-synaptic dendritic attenuations.

5. Energy-efficiency of Neuromorphic SoCs

The primary motivation for exploring NVM or RRAM-based spiking neural network is to achieve orders of magnitude improvement in energy-efficiency over the contemporary digital architectures. However, as discussed earlier, resistance range of RRAM-based synapses is a critical factor in the design of NeuSoCs. In a NeuSoC, the spike shape parameters and the *low-resistance state (LRS)* resistance, R_{LRS}, of the RRAM devices (R_{HRS} is typically order(s) of magnitude higher than R_{LRS}) contribute to the energy consumed in a spike event. The total energy consumption is also decided by the sparsity, i.e. the percentage of synapses in LRS state, spiking activity, and the power consumption in the CMOS neurons. Assuming a rectangular spike pulse-shape of amplitude A^+ and width τ^+ during the inference mode, the current input signal is $I_{syn} = \frac{A^+}{R_M}$, and the energy consumption for a spike driving a synapse with resistance R_M, $\frac{R_{LRS}}{M} < R_M < \frac{R_{HRS}}{M}$, is given by $E_{spk} = \frac{A^{+2}\tau^+}{R_M} < \frac{A^{+2}\tau^+}{R_{LRS}}$. Here in this calculation, compound synapses with $M = 16$ RRAMs in parallel are employed to achieve 4-bit resolution.

The approximate SNN energy consumption for one event can be formulated as

$$E_{SNN} = \eta_{sp}\eta_{LRS}N_sE_{spk} + N_nP_n\tau^+ \quad (1)$$

where $\eta_{sparsity}$ is the sparsity factor (i.e. the fraction of neurons firing on average), η_{LRS} is the fraction of synapses in the LRS-state, N_s is the number of synaptic connections, N_n is the number of neurons. P_{neuron} is the neuron power consumption; energy consumed in the peripheral circuits is ignored to simplify the analysis. To provide a rough system-level comparison, the AlexNet convolutional neural network for deep learning used for the Imagenet Challenge comprised of 61 million synapses and 640k neurons [53]. We assume that an equivalent SNN is constructed through transfer learning [42], or spike-based equivalent of backpropagation algorithm [46]; the circuit architecture is essentially the same. With an estimation based on the RRAM-compatible spiking neuron chip realized in [38], 4-bit compound memristive synapses [32,52,54], 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 [55] (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 [38,56]. This

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 July 2018 doi:10.20944/preprints201807.0362.v1

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}	$100k\Omega$	$1M\Omega$	$10M\Omega$
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}	$422.6\mu J$	$42.33\mu J$	$4.24\mu J$
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. Conclusions

In summary, not only the neuromorphic deep learning architectures will provide a new growth pathway for computing beyond the Moore's scaling of transistors and the variability-inflicted von Neumann computers, but also place general purpose Artificial Intelligence in the hands of end consumers, instead confining it to energy-intensive computing clusters. This can provide a breakaway for memory technology development, where memory itself can be the next generation platform and integral to computing. In-memory computation occurring in NeuSoC architectures will

place the emphasis on dense integration of memory arrays with peripheral neural circuits, extending to 3D stacking and networks of-and-on chips. 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 NSF CAREER award EECS-1454411 and Micron Foundation for the endowment. The authors also thank Prof. Maria Mitkova for discussions on experimental RRAM characterization, and Prof. John Chiasson for technical discussions on SNNs.

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. wrote 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.S.; DeBenedictis, E.P. OSTP Nanotechnology-Inspired Grand Challenge: Sensible Machines, 2015.
2. Hinton, G.E.; Salakhutdinov, R.R. Reducing the dimensionality of data with neural networks. *Science* **2006**, *313*, 504–507.
3. Nervana. Making Machines Smarter.
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 **2007**.
7. Nessler, B.; Pfeiffer, M.; Buesing, L.; Maass, W. Bayesian computation emerges in generic cortical microcircuits through spike-timing-dependent plasticity **2013**.
8. Merolla, P.A.; Arthur, J.V.; Alverez-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. Conf. Proc. IEEE Eng. Med. Biol. Soc, 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. 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.
17. 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*.
18. Yang, J.J.; Strukov, D.B.; Stewart, D.R. Memristive devices for computing. *Nature nanotechnology* **2013**, *8*, 13–24.
19. Chang, T.; Yang, Y.; Lu, W. Building neuromorphic circuits with memristive devices. *Circuits and Systems Magazine, IEEE* **2013**, *13*, 56–73.
20. Yu, S.; Kuzum, D.; Wong, H.S.P. Design considerations of synaptic device for neuromorphic computing. Circuits and Systems (ISCAS), 2014 IEEE International Symposium on. IEEE, 2014, pp. 1062–1065.
21. Indiveri, G.; Legenstein, R.; Deligeorgis, G.; Prodromakis, T. Integration of nanoscale memristor synapses in neuromorphic computing architectures. *Nanotechnology* **2013**, *24*, 384010.
22. 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.
23. Saxena, V. Memory Controlled Circuit System and Apparatus, 2015. US Patent App. 14/538,600.
24. Saxena, V. A Compact CMOS Memristor Emulator Circuit and its Applications. *arXiv preprint arXiv:1711.06819* **2017**.
25. Saxena, V.; Wu, X.; Zhu, K. Energy-Efficient CMOS Memristive Synapses for Mixed-Signal Neuromorphic System-on-a-Chip. Circuits and Systems (ISCAS), 2018 IEEE International Symposium on. IEEE, 2018, pp. 1–5.
26. 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. Electron Devices Meeting (IEDM), 2011 IEEE International. IEEE, 2011, pp. 31–6.
27. 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. Electron Devices Meeting (IEDM), 2012 IEEE International. IEEE, 2012, pp. 20–3.
28. Kozicki, M.N.; Mitkova, M.; Valov, I. Electrochemical Metallization Memories. In *Resistive Switching*; Wiley-Blackwell, 2016; pp. 483–514.
29. 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.
30. Micron. 3D XPoint? Technology: Breakthrough Nonvolatile Memory Technology.
31. Strukov, D.B.; Snider, G.S.; Stewart, D.R.; Williams, R.S. The missing memristor found. *nature* **2008**, *453*, 80.
32. 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.
33. 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.
34. 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.
35. Koch, C. Computation and the single neuron. *Nature* **1997**, *385*, 207.
36. Wu, X.; Saxena, V. Energy-efficient CMOS Neurons for Crosspoint Synapses. *submitted to the IET Electronics Letters* **2014**.

37. Wu, X.; Saxena, V.; Zhu, K. A CMOS spiking neuron for dense memristor-synapse connectivity for brain-inspired computing. Neural Networks (IJCNN), 2015 International Joint Conference on. IEEE, 2015, pp. 1–6.
38. 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.
39. Latif, M.R. Nano-Ionic Redox Resistive RAM–Device Performance Enhancement Through Materials Engineering, Characterization and Electrical Testing **2014**.
40. 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.
41. Diehl, P.U.; Cook, M. Unsupervised learning of digit recognition using spike-timing-dependent plasticity. *Frontiers in computational neuroscience* **2015**, *9*.
42. 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.
43. Lee, J.H.; Delbruck, T.; Pfeiffer, M. Training deep spiking neural networks using backpropagation. *Frontiers in Neuroscience* **2016**, *10*.
44. Kheradpisheh, S.R.; Ganjtabesh, M.; Thorpe, S.J.; Masquelier, T. STDP-based spiking deep neural networks for object recognition. *arXiv preprint arXiv:1611.01421* **2016**.
45. 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**.
46. Neftci, E.; Augustine, C.; Paul, S.; Detorakis, G. Event-driven Random Back-Propagation: Enabling Neuromorphic Deep Learning Machines. *arXiv preprint arXiv:1612.05596* **2016**.
47. 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. Electron Devices Meeting (IEDM), 2015 IEEE International. IEEE, 2015, pp. 17–1.
48. 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.
49. 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.
50. Pedrosa, V.; Clopath, C. The Role of Neuromodulators in Cortical Plasticity. A Computational Perspective. *Frontiers in synaptic neuroscience* **2016**, *8*.
51. 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**.
52. 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.
53. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 2012, pp. 1097–1105.
54. Bill, J.; Legenstein, R. A compound memristive synapse model for statistical learning through STDP in spiking neural networks. *Frontiers in neuroscience* **2014**, *8*.
55. Nvidia. New Pascal GPUs Accelerate Inference in the Data Center, 2016.
56. Saxena, V.; Baker, R.J. Indirect compensation techniques for three-stage CMOS op-amps. Circuits and Systems, 2009. MWSCAS'09. 52nd IEEE International Midwest Symposium on. IEEE, 2009, pp. 9–12.