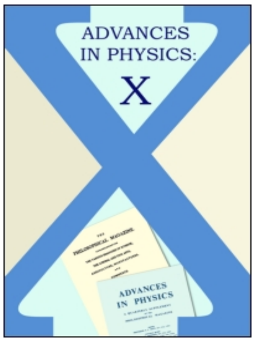


Taylor & Francis
Taylor & Francis Group

Advances in Physics: X

ISSN: (Print) 2374-6149 (Online) Journal homepage: https://www.tandfonline.com/loi/tapx20

Neuromorphic computing using non-volatile memory

Geoffrey W. Burr, Robert M. Shelby, Abu Sebastian, Sangbum Kim, Seyoung Kim, Severin Sidler, Kumar Virwani, Masatoshi Ishii, Pritish Narayanan, Alessandro Fumarola, Lucas L. Sanches, Irem Boybat, Manuel Le Gallo, Kibong Moon, Jiyoo Woo, Hyunsang Hwang & Yusuf Leblebici

To cite this article: Geoffrey W. Burr, Robert M. Shelby, Abu Sebastian, Sangbum Kim, Seyoung Kim, Severin Sidler, Kumar Virwani, Masatoshi Ishii, Pritish Narayanan, Alessandro Fumarola, Lucas L. Sanches, Irem Boybat, Manuel Le Gallo, Kibong Moon, Jiyoo Woo, Hyunsang Hwang & Yusuf Leblebici (2017) Neuromorphic computing using non-volatile memory, Advances in Physics: X, 2:1, 89-124, DOI: 10.1080/23746149.2016.1259585

To link to this article: https://doi.org/10.1080/23746149.2016.1259585

Published online: 04 Dec 2016.

ADVANCES IN PHYSICS: X, 2017
VOL. 2, NO. 1, 89–124
http://dx.doi.org/10.1080/23746149.2016.1259585

Taylor & Francis
Taylor & Francis Group

REVIEW ARTICLE

OPEN ACCESS

Neuromorphic computing using non-volatile memory

Geoffrey W. Burr[a], Robert M. Shelby[a], Abu Sebastian[b], Sangbum Kim[c], Seyoung Kim[c], Severin Sidler[d], Kumar Virwani[a], Masatoshi Ishii[e], Pritish Narayanan[a], Alessandro Fumarola[a], Lucas L. Sanches[a], Irem Boybat[b], Manuel Le Gallo[b], Kibong Moon[f], Jiyoo Woo[f], Hyunsang Hwang[f] and Yusuf Leblebici[d]

[a]IBM Research - Almaden, San Jose, CA, USA; [b]IBM Research - Zurich, Rüschlikon, Switzerland; [c]IBM T. J. Watson Research Center, Yorktown Heights, NY, USA; [d]EPFL, Lausanne, Switzerland; [e]IBM Tokyo Research Laboratory, Tokyo, Japan; [f]Department of Material Science and Engineering, Pohang University of Science and Technology, Pohang, Korea

ABSTRACT

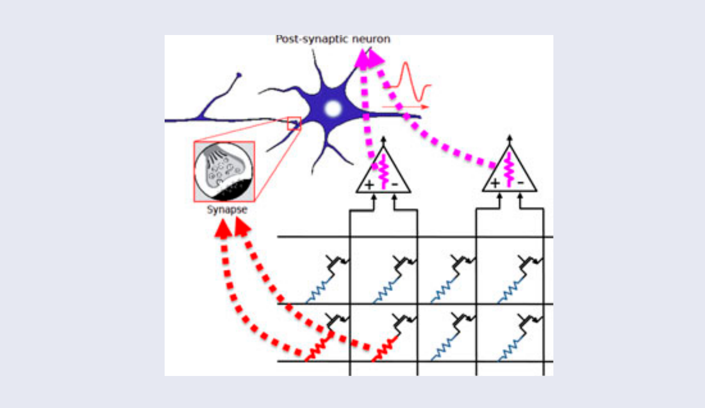
Dense crossbar arrays of non-volatile memory (NVM) devices represent one possible path for implementing massively-parallel and highly energy-efficient neuromorphic computing systems. We first review recent advances in the application of NVM devices to three computing paradigms: spiking neural networks (SNNs), deep neural networks (DNNs), and ‘Memcomputing’. In SNNs, NVM synaptic connections are updated by a local learning rule such as spike-timing-dependent-plasticity, a computational approach directly inspired by biology. For DNNs, NVM arrays can represent matrices of synaptic weights, implementing the matrix–vector multiplication needed for algorithms such as backpropagation in an analog yet massively-parallel fashion. This approach could provide significant improvements in power and speed compared to GPU-based DNN training, for applications of commercial significance. We then survey recent research in which different types of NVM devices – including phase change memory, conductive-bridging RAM, filamentary and non-filamentary RRAM, and other NVMs – have been proposed, either as a synapse or as a neuron, for use within a neuromorphic computing application. The relevant virtues and limitations of these devices are assessed, in terms of properties such as conductance dynamic range, (non)linearity and (a)symmetry of conductance response, retention, endurance, required switching power, and device variability.

ARTICLE HISTORY
Received 31 August 2016
Accepted 1 November 2016

KEYWORDS
Neuromorphic computing; non-volatile memory; spiking neural networks; spike-timing-dependent-plasticity; vector–matrix multiplication; NVM-based synapses; NVM-based neurons

PACS
(87.85.dq) Neural networks; (87.19.ll) models of single neurons and networks; (87.19.lv) learning and memory; (87.19.lw) plasticity

CONTACT Geoffrey W. Burr ✉ gwburr@us.ibm.com



1. Introduction

For more than five decades, the flexibility of the 'stored program' Von Neumann (VN) architecture has driven exponential improvements in system performance. However, as device scaling has slowed due to power- and voltage-considerations, the time and energy spent transporting data across the so-called 'Von-Neumann bottleneck' between memory and processor has become problematic. This is particularly true for data-centric applications, such as real-time image recognition and natural language processing, where state-of-the-art VN systems work hard to match the performance of an average human.

The human brain suggests an intriguing Non-Von Neumann (Non-VN) computing paradigm for future computing systems. Characterized by its massively parallel architecture connecting myriad low-power computing elements (neurons) and adaptive memory elements (synapses), the brain can outperform modern processors on many tasks involving unstructured data classification and pattern recognition. The scaling of dense non-volatile memory (NVM) crossbar arrays to few-nanometer critical dimensions has been recognized as one path to build computing systems that can mimic the massive parallelism and low-power operation found in the human brain [1–9]. The human brain has a high degree of connectivity, with any given neuron having as many as 10,000 inputs from other neurons. Dense arrays of NVM elements provide an opportunity to emulate this connectivity in hardware, if various engineering difficulties can be overcome.

These challenges include the need for a robust computational scheme, the need for peripheral circuitry that can support massively parallel access to NVM arrays, the need for an integrated crossbar 'selection device', and the need to understand the impact of the inherent limitations of NVM devices (finite dynamic range, imperfect device reliability and variability, and the non-zero programming energy) on network performance.

Since the physical properties and switching behaviors of NVM devices vary considerably with device type, various *computational schemes* have been proposed for implementing both synapses and neurons in neuromorphic networks using NVM. In some cases, it is sufficient for a synapse to be connected or not, so a binary 'on/off' NVM response (such as those found in Conductive-Bridging RAM (CBRAM) or Magnetic RAM devices) can be useful. If analog synaptic weights are desired, an NVM with continuously variable conductance – such as Phase Change Memory (PCM) or Resistive-RAM (RRAM) – may be more suitable. In some cases, multilevel synaptic weights have been represented by a number of binary NVMs connected in parallel, using stochastic programming techniques [7,9–11].

Peripheral circuitry has been proposed for the realization of hybrid CMOS/NVM neuromorphic chips [3,12–14]. Exploiting the density of nanoscale crossbar NVM arrays requires compact, efficient CMOS neurons and/or innovative techniques for matching the disparate length scales. For instance, the 'CrossNet' concept proposed 3D integration technology to form connections between a dense synapse circuit layer and a less-dense CMOS neuron circuit layer [11,14,15].

In a crossbar array, a *selection device* [16,17] is needed in series with the NVM element to prevent parasitic 'sneak path' currents that can lead to incorrect sensing of the NVM conductance. Such a device should pass very low current at low bias and have a strong nonlinear turn-on at some threshold voltage. Ideally, both the selection function and the memory function would be incorporated into the same device [11,16]. Although neuromorphic applications can tolerate more variability than conventional storage, it is still essential that very few access devices fail as a short. Variability among access devices cannot be so large that different devices might contribute significantly different read currents (into an integrate-and-fire or other accumulation circuit) for what should have been the same NVM device state, or require significantly different write conditions than their neighbors. As a result, both the role and requirements of such access devices when applied to crossbar arrays for neuromorphic applications are almost identical to those for conventional storage applications. We refer interested readers to our recent comprehensive review of access devices [16].

NVM device issues such as power dissipation, device density, conductance dynamic range, conductance retention for long-term potentiation (LTP), and device variability have been considered, as well as techniques for implementing simple Hebbian learning through Spike-Timing-Dependent-Plasticity (STDP). The asymmetric response of most NVM devices during programming causes weight updates of one sign to be favored over the other and leads to reduced network performance. This feature has led some authors to employ two NVM per synapse with opposite weight contribution during the update step [18,19].

Real NVM devices are decidedly *non-ideal*. However, since biology manages to construct highly functional neural networks from imperfect neurons and synapses, there is reason to hope that neuromorphic systems could be similarly insensitive to the presence of defective and variable devices [4,19]. It has been argued that, given enough connectivity and degrees of freedom, the network would adjust the strength of its connections to 'rewire itself to retain performance while avoiding defective' devices [20].

Given the incredibly low power (10–15 W) of the human brain, *energy efficiency* is one of the driving forces behind neuromorphic computing. A number of studies have addressed this point [21–23], particularly by using sparse computing techniques to maximize the computing functionality produced per device programming event. NVM switching that requires only low currents (0.1–1 uA), low voltages (< 1V), and short pulses (0.1–1 us) would be highly preferred [21].

In this paper, we survey the state of neuromorphic computing using NVM. We first discuss computational schemes, starting with biologically-inspired spiking neural networks (SNN) – including those based on the local STDP learning-rule. We also briefly survey pure CMOS neuromorphic implementations that do not depend on NVM. We then discuss the neuromorphic applications of crossbars as vector–matrix multipliers, including implementation of Deep Neural Networks (DNNs) based on the backpropagation learning-rule. We also discuss non-neuromorphic 'memcomputing' applications offered by NVM device arrays.

Then we survey neuromorphic computing work by the type of NVM device employed as a synapse, including PCM, CBRAM, Filamentary RRAM, non-filamentary RRAM, and other types of devices. Finally, we discuss neuromorphic research in which the NVM device plays the role of the neuron (rather than synapse), before concluding the paper.

2. Computational schemes

2.1. Spike-timing-dependent-plasticity

In the nervous system, neurons pass electrical and chemical signals to other neurons through *synapses*. STDP is a biologically-observed process for strengthening or weakening these connections [24,25]. STDP depends on the relative timing between 'action potentials' (spikes) within the input (pre-synaptic) and output (post-synaptic) neuron. In LTP, synapses are persistently strengthened by causal spiking (pre-synaptic spike occurs before the post-synaptic spike); in Long-Term Depression (LTD), acausal spiking decreases synaptic strength (Figure 1) [26]. This synaptic plasticity, or change of synaptic weight, is believed to play a key role in how the brain implements learning and memory [27]. Thus STDP can be considered as one implementation of Hebbian learning [28] – summarized by Siegrid Löwel as 'Cells that fire together, wire together' [29].

Artificial implementations of this spike-based synaptic plasticity, using asynchronous spikes of identical amplitude and duration, are often referred to as SNN [30]. Unfortunately, hardware implementations of conventional DNN that use similar asynchronous spikes solely for low-power node-to-node communication without global clocking are sometimes also referred to as SNN. Please note that SNNs as we have defined them here – employing spikes-for-learning – also

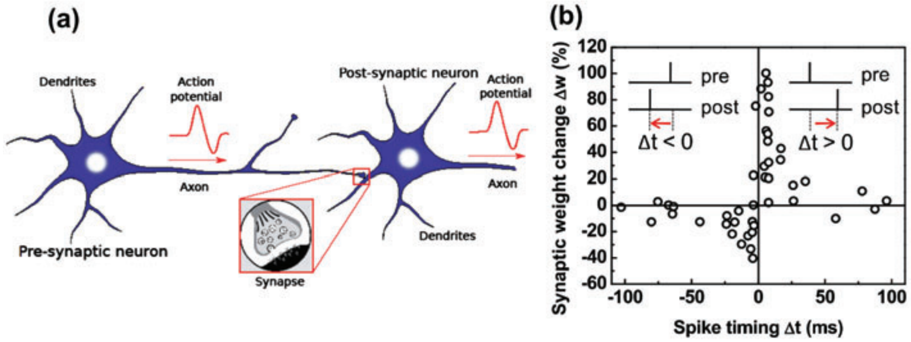


Figure 1. (a) In biology, excitatory and inhibitory postsynaptic potentials are delivered from one neuron to the next through chemical and electrical messaging at synapses, driving the generation of new 'action potentials'. (b) Biological synaptic weight change has been observed to depend on the relative timing of pre- and post-synapse spikes ($\Delta t = t_{post}$-t_{pre}), as observed in hippocampal glutamatergic synapses (after [4,26]).

provide the same energy-efficiency benefits of such spikes-for-communication approaches, if the spike occurrences are suitably sparse. In SNNs, much like the way the brain is assumed to function, information is encoded into the timing and frequency of the spikes [30]. In an SNN, spiking of the pre-synaptic neuron (electrical current) modifies the membrane potential (electrical voltage) of the post-synaptic neuron, by an amount determined by the 'synaptic weight' (electrical conductance), leading eventually to a post-synaptic spike through a leaky-integrate-and-fire or other similar neuron model [31].

The mapping of STDP as a local learning rule for NVM arrays is highly intuitive (Figure 2). One edge of the crossbar array represents pre-synaptic neurons, an orthogonal edge represents post-synaptic neurons, and the voltage on the wiring leading into these latter neurons represents membrane potential. Thus one need only implement the STDP learning rule to modify the NVM conductance based on the timing of the pulses in the pre- and post-synaptic neurons. As mentioned earlier, STDP can be implemented even with NVM devices that support small conductance change only in one direction, by separating the LTP and LTD functionality across two devices (Figure 2) [18]. That said, STDP is only a *local learning rule*, not a full computational architecture – and thus significant research still must be performed to identify applications and system architectures where STDP can be harnessed to provide robust and highly useful non-VN computation.

Despite this, a more complete understanding of STDP would help researchers understand how biological neurons use STDP within the brain. In computational neuroscience, it has been shown that many forms of biological plausible learning rule which are based on spikes bear resemblances to STDP [32–34]. Some studies have been performed in simulation, using parameters extracted from biological observations. Diehl et al. demonstrated 95% accuracy on the MNIST handwritten-digit benchmark [35] by utilizing STDP along with other

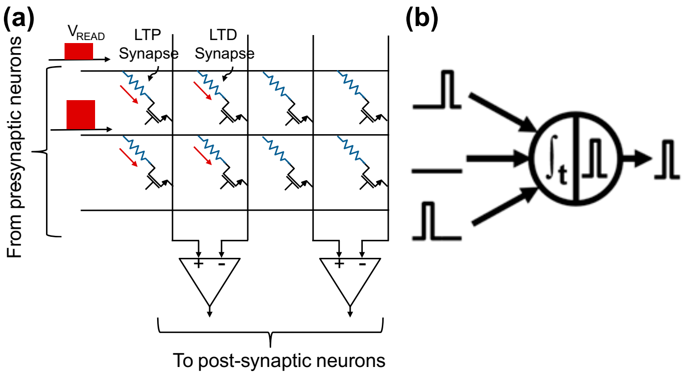


Figure 2. (a) Implementation of STDP with two-NVM-per-synapse scheme (after [18]). Due to abrupt RESET in PCM devices, LTD and LTP are implemented with SET switching in different devices, with total weight of the synapse depending on the difference between these two conductances. (b) Key spiking characteristics of spiking neural network: downstream spikes depend on the time integration of continuous inputs, with synaptic weight change dependent on relative spike timing.

biologically plausible features such as lateral inhibition and adaptive spiking threshold [36]. Masquelier et al. implemented STDP in a feed-forward multi-layer network which mimicked the ventral visual pathway, and demonstrated that this network could selectively learn salient visual features in images [37]. Later, the same group demonstrated that STDP can discover a repeating spike pattern hidden in random spike trains [38].

Even though interesting results regarding various aspects of computing functionalities of STDP have been demonstrated, each demonstration was from small networks with a limited number of modeled 'biological' neurons, using a specific set of conditions to enable the chosen computing functionality – parameters which would not necessarily work for other computing functionalities. Recent works based on SNN and STDP [39–41] have demonstrated a good inference performance on MNIST comparable to or even better in certain aspects than the conventional non-spiking implementations. However, even in these promising SNN implementations, synaptic weights were either trained in the non-spiking counterparts and later transferred to the SNN, or were trained with a specific form of STDP which had been narrowly tailored for a particular configuration.

Efficient neuromorphic hardware can be used to extend such studies, both in search of understanding the biological use of STDP as well as an architectural framework for harnessing STDP as a neuromorphic computing engine. Nearly all of this work has been performed in simulation, typically using device parameters extracted from one or a handful of NVM devices. (This implies that the statistics of how these parameters might vary across an eventual large array of such devices is either being guessed at, or worse yet, ignored.) Those studies that focus on the STDP algorithm itself are discussed in the remainder of this section; studies that focus on improving or compensating for a particular NVM 'synapse' are described in Section 3.

Unsupervised extraction of temporally correlated features was presented in [42], both for learning trajectories of a moving ball, and for detecting cars passing on a freeway based on data from a spike-based vision sensor. The authors demonstrated 98% detection accuracy after 20 min of supervised learning. The authors showed that variations in synapse and neuron parameters can significantly increase both missed car and false positive rates. Similarly, Beyeler et al. [43] discussed using STDP for MNIST image recognition, with ∼92% test accuracy with 2000 training and 1000 test epochs, although no baseline accuracy was provided. (Although a large number of SNN studies focus on MNIST and other similar image classification tasks, it is not clear, given the strong suitability of DNN for this application, whether such tasks are actually the best testbed for SNNs.)

References [44–46] implemented STDP with a generic memristive device model. In [46], the authors used ideal memristor relationships to simulate a small-scale sequence learning problem on a 25 × 25 array. Querlioz et al. studied network resilience in the presence of device variability [44,45,47]. These authors proposed homeostasis – control over the average neuron firing rate – as a potential solution for overcoming variability. An event-based simulator called 'Xnet' for modelling the impact of NVM device imperfections on STDP through semi-physical device models based on experimental characterization was also introduced [48].

In contrast to these earlier works which did not consider array design and sneak-path effects, [49] and [50] proposed a 2T1R synaptic unit cell for STDP learning. One transistor was specifically designated for integrate-and-fire, with the second transistor responsible for STDP programming. Reference [49] demonstrated a neuromorphic core with 256 neurons and 64 k PCM synaptic cells and on-line leaky-integrate-and-fire and STDP circuits for on-line and continuous training.

2.1.1. STDP with CMOS

In addition to NVM-based approaches, various CMOS-based implementations of SNNs have been demonstrated. To the researcher interested in NVM-based neuromorphic approaches, such works offer useful lessons in peripheral circuitry, computational approaches, and engineering for sparse and efficient communication within the system. Without NVM, synaptic weight is typically stored in analog or digital devices such as capacitors [51–54], 4-bit SRAM [55], 8T SRAM [56], 1-bit SRAM [57], SRAM with 8b-11b per synapse [58] and off-die SDRAM [59]. With such pure CMOS-based implementations, the reliability issues and other vagaries of emerging NVM devices can be avoided. Parameter variability is still present, especially with analog components at deep sub-micron, but is relatively well-characterized and accurate modeling is available.

When STDP [51,55] or long term plasticity [52] learning functions are implemented at each synapse (rather than neuron), parallel operation is enabled at the cost of a larger overall system size. Adding more biologically realistic functions (such as short term plasticity (STP)) further increases the size of each synapse circuit. In these various implementations, the area occupied by a single synapse circuit has varied, from 252.1 um^2 at 180 nm in [52], 100 um^2 at 180 nm in [55], and 1.6 um^2 at 45 nm node in [56]. When the number of digital synaptic bits is limited to conserve synapse circuit area, the impact of such resolution limitations must be carefully considered. The impact of hardware parameter variation on an SNN chip, Spikey, with 384 neurons and ∼100 K synapses was studied [60], as well as the impact of discretized digital weights [61].

Capacitors can store analog weight values but are leaky and subject to noise effects. DRAM or embedded DRAM is highly area-optimized for traditional memory applications, but requires frequent refreshing. SRAM is area-inefficient and subject to high leakage power. Indiveri and Liu surveyed specialized neuromorphic computing platforms including the TrueNorth [62] and NeuroGrid chips [63], and also addressed the prospects for replacing CMOS synapses with dense arrays of NVM devices [53].

Resistive-type NVM has shown a great potential in reducing the synapse circuit size compared to CMOS based implementations. The key idea is to implement STDP by designing programming pulses tailored for each NVM device and move STDP learning circuits from the synapse circuit to the neuron circuit. In addition, each NVM device can potentially replace multiple digital bits by storing analog synaptic weights. However, a system with analog weights can be less repeatable than digital weights because of increased susceptibility to noise. Implementations based on various NVM-based synapses will be discussed in Section 3.

2.2. Vector-matrix multiplication for neuromorphic applications

In contrast to SNNs, for which we have a highly biologically-realistic local-learning rule-STDP-but lack a reliable learning architecture, DNN based on the backpropagation algorithm [64] have met with tremendous recent success. Such networks, including convolutional neural networks, deep belief networks, and multilayer perceptrons, are trained using supervised learning and error backpropagation. Rather than the asynchronous, uni-valued spikes of an SNN, the neuron outputs in a DNN are real-valued numbers, processed on synchronous time-steps. While such techniques are not readily observed in biology, a serendipitous combination – objective function minimization through gradient descent implemented by backpropagation, massive labeled datasets, and the highly–parallel matrix-multiplications offered by modern GPUs – have allowed DNN to achieve considerable recent success in numerous commercially-relevant domains [65].

It has been known for some time that arrays of analog resistive memory elements are ideally suited for the multiply-accumulate operations at the heart of DNN forward-inference and training [66,67]. The multiply operation is per-

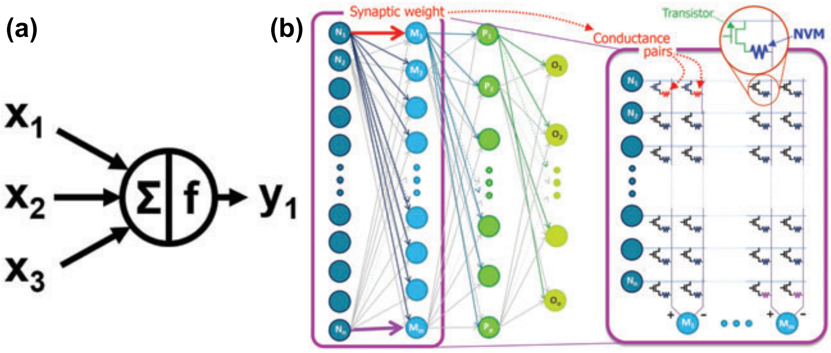


Figure 3. (a) Key characteristic of Deep Neural Networks (DNNs): downstream neuron excitation depend on the weight excitation of all upstream neurons, evaluated through a multiply-accumulate (MAC) operation. (b) This MAC operation, featured in both forward-inference and backpropagation of DNNs, can be implemented as vector–matrix multiplication on large NVM arrays. Similar to Figure 2, NVM devices with asymmetric conductance responses (e.g. RESET is more abrupt than SET, or vice versa) can still be used to train networks by simply assigning two conductances for the implementation of signed synaptic weight values.

formed at every crosspoint by Ohm's law, with current summation along rows or columns performed by Kirchhoff's current law (Figure 3). Thus these multiply-accumulate operations can be performed in parallel at the location of data with locally analog computing, reducing power by avoiding the time and energy of moving weight data [19,68–70]. Integration of device read-currents along columns implements in parallel the sums of $\Sigma \omega_{ij} x_i$ needed for forward propagation of neuron excitations x_i; integration along rows implements the sums of $\Sigma \omega_{ij} \delta_j$ needed for backpropagation of error terms δ_j [19,70].

In 2005, Senn and Fusi [10] considered the task of pattern classification in simple perceptron networks with binary synapses. Alibart et al. presented a small-scale pattern classification task using a 2 × 9 crossbar array and the delta learning rule [71]. The non-linearity of the conductance response of the NVM was observed to be a serious impediment to online learning. Other authors proposed a read-before-write scheme, with device conductance sensed before selection of the appropriate programming pulse [71–75]. While this may be quite accommodating to the peculiarities of real NVM devices, it is not clear how such an inherently-serial approach could scale to arrays with millions of NVM synapses.

Crossbar-compatible weight-update schemes were proposed [19,76–78] with upstream and downstream neurons firing programming pulses independently such that overlap of these pulses at the shared crosspoints achieves the desired weight update. Reference [19] showed no loss of test accuracy for the MNIST benchmark [35] with this scheme. Different learning rules can be implemented such as gradient descent [78,79], winner-take-all [79], and Sanger's learning rule [80], for applications such as image classification (MNIST dataset) [78], classification of cancer datasets [80], or compression and reconstruction of images [79]. The contrastive divergence at the heart of the stochastic Restricted Boltzmann Machine neural network can also be accelerated [81].

Non-linear conductance response, limited dynamic range of conductance and variability would still need to be carefully considered. References [68,82] show that some degree of non-linearity can be tolerated, so long as the conductance range over which the response is non-linear is only a small fraction (e.g. $\sim 10\%$) of the overall conductance range. The deterioration of recognition accuracy with non-linearity has also been studied for a sparse-coding algorithm [23]. With respect to dynamic range, Burr et al. [68] proposed $\sim 20 - 50$ programming steps between min and max conductance with a conductance pair representing a single weight, mirroring the conclusions of Yu et al. (6-bit resolution) [23]. Other authors have sought to quantify the impact of parameter variation and device reliability on DNN training [19,83].

Gamrat et al. applied 'spike-coding' for inference, showing competitive performance on MNIST with pre-trained weights stored on ideal devices [84]. Garbin et al. suggested that the parallel combination of 10-20 HfO_2 RRAM devices provided more robustness to device variability [85,86]. Training and forward-inference of a small one-layer DNN was implemented on a 12×12 memristor crossbar requiring no separate selection device [87,88]. Modeling based on the programming dynamics of oxide memristor devices [89] was shown to support MNIST digit classification at high accuracies [88,90]. Finally, NVM-based DNN implementations were compared to GPUs in terms of power and speed, showing the prospect for 25× speedup and orders of magnitude lower power for DNN training [69].

2.3. Non-neuromorphic applications (memcomputing)

In addition to Spiking and DNN, the physical attributes and state dynamics of emerging NVM devices can be used for computation. In this section we briefly review such 'memcomputing' approaches.

In memristive logic, also known as state-full logic, the same devices are used simultaneously to store the inputs, to perform the logic operation, and to store the output result [91]. By using a 'material implication' gate $q \leftarrow p\text{IMP}q$ (equivalent to (NOTp) ORq), Borghetti et al. showed that the NAND logic operation could be performed using three memristive devices connected to a load resistor [91], implying extension to all Boolean logic (since NAND is logically complete). Later, it was shown that such IMP logic can be implemented in a memristive crossbar by having two memristors on the same bitline but different wordlines [92]. Additional memristive logic crossbar architectures have been demonstrated based on NOR (also logically complete) [93–95].

Other adaptations of NVM device physics to logic-in-memory operations include the rich pattern dynamics exhibited by ferroelectric domain switching [96], and the physics of crystallization [97] and melting [98] of phase-change materials. The accumulation property of phase-change materials has been exploited to perform the basic arithmetic processes of addition, multiplication, division and subtraction with simultaneous storage of the result [99], and to factor numbers in parallel [100,101]. Similar results have been obtained with Ag–Ge–Se devices [102].

A particularly attractive application area is that of random number generation (RNG) for stochastic computing and cryptography [103,104], exploiting the stochasticity associated with many NVM devices. One example is the spin-torque switching in magnetic tunnel junction (MTJ), which can be used to generate sequences of random numbers [105]. The random formation, rupture and defect composition of the conductive filament(s) within metal-oxide-based resistive memory devices offers a second source of inherent randomness [106,107]. Finally, randomness in the atomic configurations of the amorphous phase created via the melt-quench process after PCM RESET operations can be harnessed for RNG [108]. Maintaining a near-unbiased RNG with equal probabilities of 0 and 1 requires careful probability tracking and adjustment of switching parameters. Recently, a new approach using coupled RRAM devices was proposed that provides unbiased RNG without the need for such kind of additional circuitry [109].

Networks of NVMs can perform certain computational tasks with remarkable efficiency. For instance, a network of resistive devices can find all possible solutions in multiple-solution mazes and sort out the solution paths according to their lengths [110]. Recent proposals for Universal Memcomputing Machine, a class of general-purpose machines based on interacting memcomputing elements for solving computationally hard problems, have initiated discussions on new physical models of computation [111–113].

3. NVM as a synapse

Desirable properties and switching dynamics of a variety of NVM devices in the context of neuromorphic applications have been discussed by several authors [2,4,8,9].

3.1. PCM as a synapse

PCM (Figure 4(a)) depends on the large difference in electrical resistivity between the amorphous (low-conductance) and crystalline (high-conductance) phases of so-called phase change materials [114–116]. In NVM devices in general, the process of programming into the high-conductance state is referred to as ‘SET,’ and programming into low-conductance is ‘RESET’. PCM can be applied to neuromorphic applications where ‘device history’ is desirable, but only the SET process can be made incremental, with repetitive pulses slowly crystallizing a high-resistance amorphous plug within the device. Since RESET involves melt and quench, it tends to be an abrupt process, especially within an array of not quite homogenous devices.

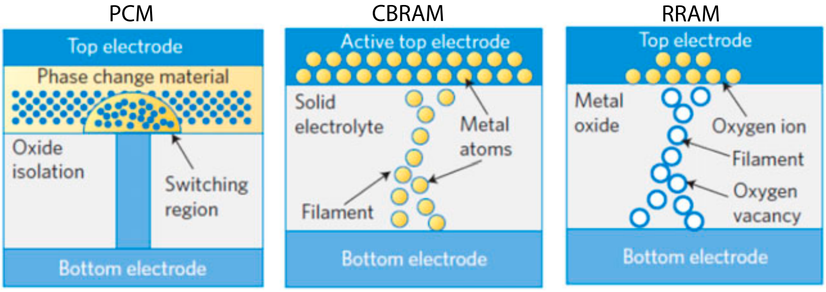


Figure 4. (a) Phase change memory (PCM) depends on the large difference in electrical resistivity between the amorphous (low-conductance) and crystalline (high-conductance) phases of so-called phase change materials [114–116]. (b) Conductive-Bridging RAM is based on the electrochemical formation of conductive metallic filaments through an insulating solid-electrolyte or oxide [117,118]. (c) The conductive filaments in a filamentary RRAM are chains of defects through an otherwise insulating thin-film oxide [119].
Note: (After [120]).

As mentioned earlier, a two-PCM approach was proposed [18,121] to implement STDP by using different devices for LTP and LTD (see Figure 2). In this implementation, when an input neuron spikes, it sends out a read pulse and enters 'LTP mode' for time t_{LTP}. If the post-synaptic neuron spikes during this period, the LTP synapse receives a partial SET pulse; otherwise, the LTD synapse is programmed.

Suri et al. improved synaptic performances of standard $Ge_2Sb_2Te_5$ (GST)-based PCM devices by an additionally introducing a thin HfO_2 layer [122,123]. Increased dynamic range was attributed to the effect of this interface layer on crystallization kinetics (through the activation energies related to growth and nucleation sites). The two-PCM approach, much like the later use of PCM arrays for vector–matrix computation of the backpropagation algorithm [19], requires a burdensome refresh protocol to deal with fully SET synapses, in which inputs are disabled, effective weights are read, and conductances RESET where necessary to maintain weights but with lower conductance values.

Suri et al. used the behavioral model (from measured device data) of GST and GeTe PCM to run a Xnet event-based simulation to extract features from a dynamic vision sensor and count cars in six highway traffic lanes [123]. In later work, a circuit-compatible model incorporating the electrical and thermal characteristics of the top and bottom electrodes together with phase change material parameters was developed [124]. Authors observed that the maximum conductance was reached in fewer pulses if either growth or nucleation rate was enhanced, growth and nucleation rates strongly influenced the shape (but not size) of the amorphous plug after RESET pulses, and conductance during partial-SET was more sensitive to nucleation rate than growth rate. Since growth-dominated GeTe saturated in conductance more quickly than nucleation-dominated GST, GST could offer more intermediate conductance states than GeTe.

Symmetric and asymmetric STDP was also implemented with a single PCM cell per synapse [125,126]. RESET pulses of varying amplitude, and staircase-down pulses of varying amplitudes were used for partial SET. The use of short pulse timings kept the total energy low despite high programming currents, and associative and sequential learning were demonstrated in simulation. Later the same authors showed that by tuning the energy of the spikes, the total energy for the neuromorphic implementations can be reduced [127]. Ab initio molecular dynamics were used to model physical changes inside phase change materials during STDP potentiation and depression with a stepwise increase in the material order in response to heat pulses (not electrical pulses) of different heights and lengths [128].

Eryilmaz et al. experimentally demonstrated array-level learning using a 10 × 10 array of transistor-selected PCM cells, showing Hebbian STDP learning of several simple patterns [129]. Higher initial resistance variation was shown to require longer training. Ambrogio et al. used measurements of a few transistor-selected PCM cells (45 nm node) to simulate larger networks [130]. With a two-layer network of 28 × 28 pre- and 1 post-neuron, MNIST digit recognition probability was 33%, with a corresponding error of 6%. With a three-layer network, recognition probability reached as high as 95.5% for 256 neurons (error of 0.35%). The authors also discussed the ability of their network to forget previous information and learn new information both in parallel and in sequence.

Li and Zhong et al. analyzed four different variants of STDP updates (anti-symmetric Hebbian and anti-Hebbian update with potentiation, and symmetric update with depression and potentiation) by applying pulses in different time windows [131]. These forms were implemented on a few devices and in simulation [132]. Jackson et al. implemented two STDP schemes: by generating STDP-encoded neuronal firing delays within the electronic pulses arriving at the synapse, and by tracking the delay using a simple RC circuit in each neuron [133]. The latter approach was shown to be achievable using programming energies less than 5 pJ in 10 nm pore (19 nm actual) phase change devices. Authors then simulated 100 leaky integrate and fire neurons to successfully learn a simple task of predicting the next item in a sequence of four stimuli.

3.2. CBRAM as a synapse

CBRAM (Figure 4(b)), based on the electrochemical formation of conductive metallic filaments [117,118], is a promising approach for future NVM device applications due to its fast speed (∼ns), scalability to the nanometer regime, and ultra-low power consumption (∼nW). One of the difficulties for neuromorphic applications is the inherently abrupt nature of the filament formation (SET) process. While filament broadening can be performed incrementally, the resulting states are quite conductive, leading to large overall currents for any neuromorphic system summing read currents across many devices. For instance, integrate-and-fire neurons would then require rather large capacitors.

Ohno et al. implemented STDP synaptic operation with a silver sulfide (Ag_2S) gap-type 'atomic switch.' [134]. Short-term memory formation depended on both pulse amplitude and width, with lower amplitude and narrower width requiring a larger number of input pulses. Application of several pulses within a short interval established a relatively stable state, interpreted as long term potentiation (LTP). A 7 × 7 array of inorganic synapses was used to experimentally demonstrate memorization (and forgetting) of two simple patterns [135].

Yu et al. developed a physical model to investigate the switching dynamics of CBRAM devices [136]. During SET (positive bias), the top electrode is oxidized, with metal ions drifting to the bottom electrode where they are reduced, producing a metallic filament that grows vertically until it reaches the top electrode. During RESET, the filament tends to dissolve laterally first due to the enhanced lateral electric field at the top of the filament. Key parameters such as the activation energy and effective hopping distance were extracted from electrical measurements. A signaling scheme was designed to implement STDP-like conductance change by tuning pulse amplitude.

A layered memristor device structure, in which the continuous motion of Ag nanoparticles co-sputtered in a Si layer produced reliable analog switching, was demonstrated [137]. Two CMOS-based integrate-and-fire neurons and a 100 nm × 100 nm memristor device were connected to demonstrate STDP, and 1.5 e8 potentiation and depression cycles were demonstrated without significant degradation. Suri et al. proposed the use of CBRAM devices as binary synapses in low-power stochastic neuromorphic systems [138]. Binary CBRAM synapses and a stochastic STDP learning rule allowed asynchronous analog data streams to be processed for recognition and extraction of repetitive, real-time auditory and visual patterns in a fully unsupervised manner.

Ziegler et al. demonstrated that a single Pt/$Ge_{0.3}Se_{0.7}$/SiO_2/Cu memristive device implemented in analog circuitry mimics non-associative and associative types of learning [139]. Yang et al. reported nanoscale and solid-state physically evolving networks based on memristive effects that enable the self-organization of Ag nanoclusters [140]. Sillin et al. developed a numerical model based on the synapse-like properties of individual atomic switches and the random nature of the network wiring [141].

3.3. Filamentary RRAM as a synapse

Filamentary RRAM (Figure 4(c)) is an NVM that is quite similar to CBRAM, except that the filament through an otherwise insulating thin-film is a chain of defects within an oxide (rather than a chain of metallic atoms of one of the two electrodes through an insulating solid-electrolyte or oxide) [119]. Filamentary RRAM is attractive because it requires only metal-oxides such as HfOx, AlOx, WOx, FeOx, GdOx, TaOx and TiOx and mixtures/laminates of such films, many of which are already in use in CMOS fabs. The multi-level or gradual memory modulation needed to imitate adaptive synaptic changes has been demonstrated in most of these materials. The underlying metal-insulator–metal structure is simple, compact, CMOS-compatibile and highly scalable, and the energy consumption per synaptic operation and programming currents can be made ultralow (sub-pJ energies, < 1 uA programming current).

However, like CBRAM, the filament formation/completion process is inherently abrupt and difficult to control. Low-power switching is attractive, but since it is typically achieved by moving only a handful of atomic defects, large variability through Poissonian statistics ('shot noise') is unavoidable. Once triggered (by electrical field and/or local temperature increases), the filament formation/completion in both filamentary RRAM and CBRAM must be truncated by the external circuit lest the filament grow too thick to be removed by any accessible RESET pulse. This is particularly problematic for neuromorphic applications, since a single highly-conductive device with a thick filament is contributing much more current into a vector sum or leaky-integrate-and-fire than its neighbors.

Choi et al. reported a gradual RESET switching with increasing RESET voltages in a GdOx-based RRAM and crossbar array with abrupt SET switching [142]. Yu et al. and Wu et al. demonstrated multi-level switching in SET operation with continuously increasing external compliance currents, and in RESET with continuously increasing reset voltages in TiN/HfOx/AlOx/Pt and TiN/Ti/AlOx/TiN RRAM devices, respectively [143,144]. In [50], STDP pulses were engineered based on characteristics of bipolar Si-doped HfO_2 memory cell, and tested on individual devices. Simulations were performed to demonstrate learning and recognition of small black and white patterns. Chua and co-workers proposed memristive cellular automata networks [145] and a memristor bridge circuit intended to simplify the 'chip-in-the-loop' training of DNN [146,147].

Resistance modulation by controlling current compliance and/or pulse voltages has numerous disadvantages in terms of peripheral circuit design and complexity. Similar to approaches taken to avoid PCM SET, Yu et al. proposed the use of only the RESET operation in Filamentary RRAM, achieving hundreds of resistance states using a forming-free Pt/HfOx/TiOx/HfOx/TiOx/TiN multilayer oxide-based synaptic device [148,149]. Short pulses (10 ns) with an identical pulse amplitude enabled sub-pJ energy per spike with potentially simple neuron circuits. A stochastic compact model was developed to quantify the gradual resistance modulation and was applied to a large-scale artificial visual system simulation. 1024 neuron circuits and 16,348 oxide-based synapses were modeled to simulate a 2-layer neural network, showing tolerance to resistance variations of up to 10%.

Another approach is to embrace the abrupt SET operation and adopt binary stochastic switching synapses [150]. A two-layer winner-take-all neural network (4096 synapses connecting 32 × 32 input neurons to 2 × 2 output neurons) were simulated for an orientation classification task. Jeong et al. proposed the use of a long and low-amplitude heating pulse in addition to regular SET pulses, to improve the analog switching of TaOx-based RRAM devices [151]. The heating pulse pre-heats the filament by local Joule heating, enhancing diffusion during short SET pulses and enhancing SET dynamic range by 80%.

The impact of a multi-step 'forming' process (for generation of the first filament in the device) was studied in AlOx-based RRAM devices [152]. Wang et al. investigated the relation between compliance current during forming process and conductance stability and synaptic behavior of a FeOx-based RRAM [153]. To incorporate both voltage-controlled RESET and current-controlled SET, Ambrogio et al. proposed a 1T1R synapse with HfO_2-based RRAM [154]. The transistor serves as both selector and voltage-independent current source, improving control over filament formation during forming at SET operations. STDP modulation of the RRAM resistance was demonstrated in simulations with a compact RRAM model, and also experimentally using a transistor-selected HfO_2-RRAM with timing in the range of a few tens of ms. A 40 × 40 memristor crossbar array was integrated with CMOS circuits, and resistive switching demonstrated using the intrinsic rectifying IV characteristics of a SiGe/W/a-Si/Ag RRAM device [155]. Multi-level storage capability was verified by programming cells with different series resistances and external diodes.

Bill et al. reported the computational implications of synaptic plasticity through stochastic filament formation in multiple binary RRAM cells tied together in parallel to form a single compound stochastic synapse [156]. A HfO_2-based vertical RRAM (VRRAM) technology was recently reported by Piccolboni et al. [157]. Each synapse is composed of stack of VRRAM devices (each exhibiting only two distinct resistance states) with one common select transistor, exhibiting analog conductance behavior through the parallel configuration of N RRAM cells. Real-time auditory visual pattern extraction applications were shown via simulation. Although here there is a unique pair of electrodes for every device participating in the synaptic behavior, some researchers have proposed multiple devices in parallel between a single shared pair of electrodes. However, this cannot work because the existence of a filament in any of the devices shorts out all the parallel devices, preventing further filament formation.

Prezioso experimentally characterized Al_2O_3/TiO_2-based memristors to model the impact of conductance-dependent conductance change on STDP [158]. Simulations showed that memristors exhibiting such conductance responses enable self-adaptation of the synaptic weights to a narrow interval in the intermediate value of their dynamic range, at least in a simple spiking network. Thus suggests that non-ideal device characteristics could potentially be compensated at the system level. Cruz-Albrecht et al. designed and simulated a neuromorphic system combining a reconfigurable front-end analog processing core and W/WOx/Pd RRAM-based synaptic storage [159]. Expected power consumption for ∼70,000 nanoscale RRAMs and the 16 million CMOS transistors was estimated to be 130 mW. Deng et al. combined the internal dynamics of a TiW/Pt/FeOx/Pt/TiW RRAM with a mixed-analog–digital system (ADC block at dendritic inputs, DAC block before the crossbar RRAM) to implement a recurrent neural network using the recursive least-squares algorithm [160]. Two parallel RRAM were combined to realize both excitatory and inhibitory synapses. This neuromorphic system was shown to offer good tolerance to the variation of the RRAM devices.

3.4. Non-filamentary RRAM as a synapse

In a non-filamentary RRAM, defect migration takes place over the entire device area, typically at an interface between two materials such as an oxide and metal that form a Schottky barrier [161]. Motion of defects (dopants) towards the electrode collapses the Schottky barrier, resulting in significant resistance changes. In addition, ionic motion can transition oxide region from a highly resistive state to a more conductive state, narrowing the tunneling gap [162] and enabling gradual switching that is highly suitable for implementing an analog synapse. While this eliminates the problems associated with RRAM filaments, non-filamentary RRAMs tend to require an unpleasant tradeoff between programming speed (requiring a low energy barrier to defect diffusion) and retention (calling for a high energy barrier for this same diffusion). Critical to the use of such non-filamentary RRAM will be a viable solution to this so-called 'voltage–time dilemma' [161].

Seo et al. demonstrated analog memory enabling synaptic plasticity and STDP in a nanoscale TiOx/TiOy bilayer RRAM system, with multilevel conductance states caused by the movement of oxygen ions between the TiOy layer and the TiOx layer [163]. When 100 successive identical pulses for potentiation or depression were applied to the device, the conductance was progressively and continuously increased (decreased). Chang et al. reported a Pd/WOx/W non-filamentary memristive device fabricated by rapid thermal annealing of a tungsten film at 400°C in O_2 environment, exhibiting gradual resistance modulation characteristics attributed to uniform migration of oxygen vacancies under bias, modulating the Schottky barrier emission and tunneling at the WOx/electrode interface [164].

Similar physics was reported for a non-filamentary RRAM device based on a Ta/TaOx/TiO_2/Ti stack [165,166]. A comprehensive analytical model based on barrier modulation induced by oxygen ion migration was proposed to explain the observed synaptic resistance change characteristics in the device, including STDP and paired-pulse facilitation. Resistance states could only be read out at negative voltages due to significant non-linearity in the device characteristics. Energy consumption was as low as 7 fJ per training pulse for depression, although the switching voltages were large with long duration (+9.2 V/1 ms and –8.0 V/50 us for potentiation and depression, respectively). However, these characteristics might improve as devices are scaled down in size from the 10^4um^2 regime to the nanoscale regime, since non-filamentary devices do change significantly with critical dimension.

Resistance instability in RRAM synapses can be exploited to emulate Short-Term Modulation (STM). Lim et al. investigated the STM effect in TiO_2 material

[167] utilizing a non-filamentary switching RRAM device. This potentiation behavior depends on the stimulation frequency, with synaptic weight increasing as frequency increased. Yang et al. reported two different switching modes, volatile and nonvolatile resistance switching behaviors, in a Pt/WO_{3-x}/Pt based non-filamentary memristor device [168]. Before the device is formed, the device exhibits a volatile switching characteristics; after forming at 6 V, the device shows less volatile resistive memory-like behavior with ∼22–25 s resistance state lifetimes, potentially enabling short- and long-term memory behaviors

Park et al. demonstrated a neuromorphic system with a 1 kbit cross-point array of interface-type RRAM based on reactive metal/$Pr_{0.7}Ca_{0.3}MnO_3$ (PCMO) stacks, with device diameters ranging from 150 nm–1 um. When Al metal is deposited, the chemical reaction between Al and PCMO forms a very thin AlOx layer. When negative bias is applied to the Al layer, oxygen ions move from the AlOx to the PCMO bulk, leading to a low resistance state. Under positive bias, oxygen ions were attracted from the PCMO layer, forming a thick insulating oxide layer and a high resistance state. Continuously increasing potentiation (depression) behaviors were observed upon iterative programming with identical pulses of negative (positive) bias [169,170].

However, it was difficult to emulate gradual LTD using identical spikes because of large differences between the SET and RESET characteristics, caused by different oxidation and reduction energies. To obtain a symmetric and gradual behavior on the Al/PCMO device, a programming scheme was introduced [171]. Varying sputter conditions during the deposition of the TiNx layer was shown to improve device characteristics such as current level and on/off ratio by engineering the Schottky barrier between TiN and PCMO [172].

Sheri et al. proposed a solution for the asymmetrical memristor behavior of TiN/PCMO synapses, achieving improved pattern recognition accuracy [173]. To avoid the abrupt conductance change observed in depression mode, they proposed a two-PCMO-memristor device model in which two PCMO devices constitute one synapse (similar to earlier approaches with PCM and RRAM [18,19,121,148,149]). Moon et al. developed neuromorphic hardware for real-time associative memory characteristics, combining CMOS neurons with TiN/PCMO synapses [174]. They designed a circuit consisting of elements including an adder, a divider, and a comparator to realize classical conditioning (learning to associate ‘fear’ or ‘hunger’ with particular stimuli). To investigate the role of the electrode on PCMO-based devices, Lee et al. fabricated and analyzed the effects of various reactive top electrodes (Al, Ta, TiN, and Ni) [175]. Devices with low metal-oxide free energies exhibited an increased number of conductance states, but more asymmetric conductance change, posing a tradeoff where these effects need to be balanced for best neural network performance.

Recently, this group reported a high-density cross-point synapse array using 200 mm wafer-scale PCMO-based memristive synapses and exhibiting improved synaptic characteristics [176]. Experimental results were reported in which the memristive HNN system is used to recognize human thought patterns for three vowels: a, i, and u from electroencephalography signals. To increase the recognition accuracy, post-neurons were paired into three groups, with each group comparing two of the three vowels (a vs. i, a vs. u, and i vs. u). Fear-conditioning experiments were also performed using signals from the brain of a live rat (dentate gyrus in hippocampus) [177].

Hansen et al. demonstrated a double-barrier memristive device that consists of an ultra-thin memristive layer (Nb_xO_y) sandwiched between an Al_2O_3 tunnel barrier and a Schottky-like contact [178]. They showed that the resistive switching originates from oxygen diffusion and modification of the local electronic interface states within the Nb_xO_y layer, which then influence the interface properties of the gold contact and the Al_2O_3 tunneling barrier. They proposed the use of such devices in neuromorphic mixed signal circuits, similar to earlier experiments using TiO_2 [179].

3.5. Other types of non-volatile and partially-volatile synapses

Besides the NVM devices described so far, other devices, such as organic nanoparticle transistors [180], inorganic devices [135,181], spin devices [182–186], carbon nanotubes (CNT) [187–189], ferroelectric [190], Mott [191] and tunnel junction based [192] memristors have been used in neuromorphic applications. Using these devices as artificial synapses, several characteristics of biological synapses – such as STP [135,180,181,188], [135] LTP, LTD and STDP [184, 185,189,190,192] – have been mimicked. While some of the devices have only been fabricated, characterized and simulated [180,181,185,188], others have been used in system level simulations [182,184,186,187,189,191] or even hardware experiments [135,190,192] to show neuromorphic computing capabilities.

3.5.1. Artificial synapses based on spin devices

In [185], a magnetic heterostructure between a magnetic material exhibiting Perpendicular Magnetic Anisotropy and a non-magnetic heavy metal with high spin–orbit coupling was used to enable STDP. Device conductance was a linear function of the domain-wall position, modulated by a current flowing parallel to the domain-wall magnet (through spin–orbit torque generated by spin-polarized electrons accumulated at the interface). Pre-, input- and output-spikes were applied by access transistors connected to the heterostructure. Vincent et al. [184] used spin-transfer torque magnetic memory to implement a stochastic memristive synapse, by carefully choosing programming pulses (current and time) to implement controlled switching probabilities. A simplified STDP learning rule was applied, together with lateral inhibition to implement a winner-takes-all architecture. The detection of cars recorded on a neuromorphic retina was simulated, and the impact of device variations (minimum/maximum resistance, transient effects) tested through Monte Carlo simulations. Zhang et a. [193] proposed an all-spin DNN based on a compound spintronic synapse and neuron composed of multiple vertically stacked Magnetic Tunnel Junctions (MTJs), implementing multiple resistance states and a multi-step transfer function.

3.5.2. Artificial synapses based on CNT

In [189], the intrinsic hysteretic behavior of CNT channel conductance was used to store synaptic weights. A hysteresis loop was observed when the gate voltage is varied under constant V_{ds}, and attributed to charge trapping in water molecules in the ambient environment around the CNT. Using sawtooth-shaped pre- and post-spikes, an STDP weight update rule was implemented on a single device. System level simulations demonstrated unsupervised learning on the MNIST dataset.

In [187] an optically gated three-terminal device based on CNT coated with a photo-conducting polymer was used as an artificial synapse. A light pulse reset the channel conductance to its maximum value, with photo-induced electrons generated in the polymer layer and trapped in the gate dielectric. With its threshold voltage thus reduced, the device remained in the on-state in the dark. Device conductance could then be decreased by applying either negative bias pulses on the gate or positive bias pulses on the drain. Using an FPGA hardware setup with 8 OG-CNTFET devices, a supervised learning algorithm (based on Widrow-Hoff's least square 'Delta' rule) was used to demonstrate learning of 2- and 3-input binary, linearly separable functions. Good tolerance to device level variability (ON-state current, threshold voltage) was observed.

Shen et al. [188] built a *p*-type CNTFET, with a random CNT network as the channel and an aluminium gate on aluminium oxide implanted with indium ions. Positive voltage on the gate trapped electrons in dielectric defects, temporarily increasing hole concentration in the channel and hence its conductance. An input neuron was connected to the gate electrode, an output neuron to the source, with the drain connected to a fixed potential (negative for inhibitory, positive for excitatory device). An excitatory (inhibitory) input spike caused a sudden increase in excitatory (inhibitory) postsynaptic current through the device. The STP accumulative behavior of various input spike combinations was studied.

3.5.3. Other synaptic devices

Zhu et al. [181] suggested an in-plane lateral-coupled oxide-based artificial synapse network, consisting of an indium-zinc-oxide (IZO)-based gate with source and drain electrodes and a P-doped nanogranular SiO_2 channel. Lateral modulation is caused by proton-related electrical-double-layer effect. The gate was used as presynaptic terminal, with channel conductance encoding weight (V_{ds} constant). A positive voltage pulse on the gate induced proton migration, increasing channel conductance. Several pulses produced an accumulated EPSC, which can be interpreted as STP. Neural networks were constructed by laterally coupling several presynaptic gate terminals to the same channel, with synaptic weights spatiotemporally dependent on the spikes applied to the presynaptic input terminals.

Alibart et al. [180] proposed an artificial synapse based on a nanoparticle organic memory FET (NOMFET) using Ag nano-particles and pentacene as the organic semiconductor. The nano-particles stored charge at every pulse on the gate but released them slowly. Charge accumulation through several pulses can be interpreted as STP. Since hole-trapping led to increased channel conductance, initially charging the NPs with negative (positive) bias voltage exhibit synaptic depression (potentiation). By connecting gate to an input neuron and drain to an output neuron, high (low) frequency pulse trains were shown to exhibit conductance depression (potentiation). The working frequency range could be adapted by adjusting device and NP size.

Tunnel junction based memristors have been proposed as artificial synapses in [192]. Several materials were used: Magnesia based tunnel junctions (MgO), Barium Titanate junctions (BTO), and Tantalum Oxide (Ta-O). The conductance change is based on the effective tunnel barrier thickness, and exhibited bipolar switching with a programming voltage threshold (below which the devices were not programmed). With particular pulsing sequences, conductance could be gradually increased or decreased for LTP and LTD.

In [190], 3T-FeMEM ferroelectric memristors were used to demonstrate on-chip pattern recognition. An inverted staggered ferroelectric thin-film transistor structure and CMOS circuit exhibited hysteretic $V_{GS} - I_D$ characteristics. To demonstrate STDP learning, timing differences were converted to various pulse amplitudes (bipolar saw-tooth signal for pre-spike, rectangular pulse for post-spike). A CMOS selector enabled the overlap of pre- and post-spike to be applied at the gate of the 3T-FeMEM, creating STDP characteristics. Using 9 CMOS neurons and 16 synapses (each composed one excitatory and one inhibitory 3T-FeMEM device), a small recurrent Hopfield network was fabricated and used to demonstrate associative learning and partial recall of two 3×3 patterns.

4. NVM as a neuron

In biological neurons, a thin lipid-bilayer membrane separates the electrical charge inside the cell from that outside it, which, in conjunction with several electrochemical mechanisms, allows an equilibrium membrane potential to be maintained. With the arrival of excitatory and inhibitory postsynaptic potentials through the dendrites of the neuron, the membrane potential changes and upon sufficient excitation, an action potential is generated (‘neuronal firing’, see Figure 1) [195]. Emulation of these neuronal dynamics, including maintenance of the equilibrium potential, the transient dynamics and the process of neurotransmission, is thought to be the key to the realization of biologically plausible neuromorphic computing systems [196]. The complex neuronal dynamics captured in the Hodgkin–Huxley model and in various threshold-based neuronal models must often be simplified for efficient hardware realizations [197]. In doing so, the integration of the postsynaptic potentials (related to the neuronal

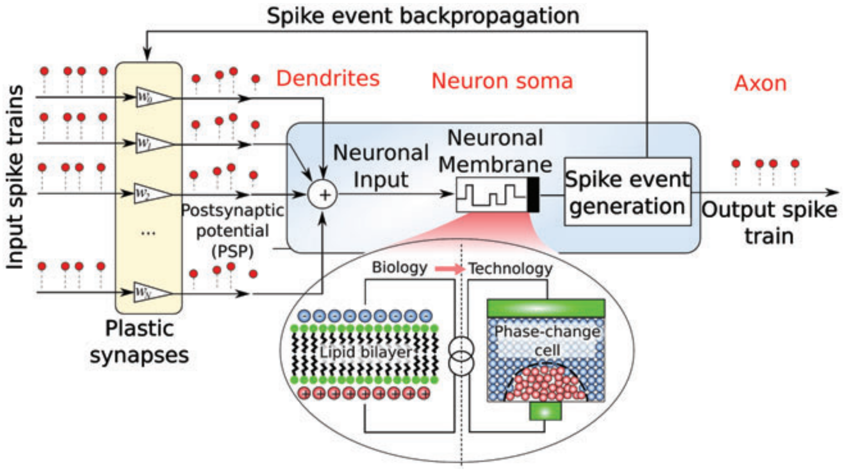


Figure 5. The physics of Non-Volatile Memory (NVM) devices can be used to emulate the dynamics of biological neurons, in order to gain significant areal/power efficiency and seamless integration with dense synaptic arrays.
Note: (After [194]).

soma) and the subsequent firing events (related to the axon) are the two most important dynamical components retained. There is a growing body of work that tries to exploit the physics of NVM devices to emulate these neuronal dynamics (Figure 5). The objective is to gain significant areal/power efficiency as well as to achieve seamless integration with dense synaptic arrays.

When an NVM is used as a neuron, it is not essential that it achieves a continuum of conductance states, but instead that it implements an accumulative behavior, 'firing' after a certain number of pulses have been received. These pulses could be changing an internal state which is not necessarily well reflected in the external conductance until the threshold is reached and the neuron 'fires.' In contrast, each conductance state of an NVM used as a synapse is important, because it impacts the network through how it contributes into an integrate-and-fire or other accumulation operation. Furthermore, for NVMs used as a neuron, non-volatility is not essential; volatility could potentially be used to implement leaky integrate-and-fire dynamics.

Ovshinksy and Wright first suggested the use of PCM devices for neuronal realization [99,198]. Tuma et al. recently showed that post-synaptic inputs can be integrated using PCM-based neurons [194], experimentally demonstrating the evolution of neuronal membrane potential as encoded by phase configuration within the device. The detection of temporal correlations within a large number of event-based data streams was demonstrated. All-PCM neuromorphic systems, in which both the neuronal and the synaptic elements were realized using phase-change devices, have also been reported [199,200].

Recent studies show that in addition to the deterministic neuronal dynamics, stochastic neuronal dynamics also play a key role in signal encoding and transmission – for example, in biological neuronal populations that represent and transmit sensory and motor signals. This stochastic behavior can be attributed to a number of complex phenomena, such as ionic conductance noise, chaotic motion of charge carriers due to thermal noise, inter-neuron morphologic variabilities, and other background noise [201]. Emulating this stochastic firing behavior within artificial neurons could enable intriguing functionality [202].

Mimicking this stochastic neuronal behavior at the device level, Tuma et al. showed that neuronal realizations using PCM devices exhibit significant interneuronal as well as intra-neuronal randomness [194]. Intra-device stochasticity – attributed to shot-to-shot variability in both thickness and internal atomic configuration of the melt-quenched amorphous region – causes multiple integrate-and-fire cycles in a single phase-change neuron to produce a distribution of the interspike intervals, enabling population-based computation. Fast signals were demonstrated to be accurately represented by overall neuron population despite the 'too-slow' firing rate of the individual neurons.

However, it should be noted here that the melt-quenching process within PCM devices, and in particular elemental migration within the molten state, is the step that tends to limit device endurance. Similarly in RRAM devices, large changes in conductance tend to 'use up' more endurance than the smaller changes in conductance involved in synaptic plasticity. Thus, if each neuron will only be able to fire a finite number of spikes, it will be important to ensure that spike firing is extremely sparse across the lifetime of the system. Alternatively, other material systems offering higher endurance could be considered.

Al-Shedivat et al. proposed stochastic artificial neurons using TiO_x-based resistive memory devices [203]. Integration of neuronal inputs leads to a large voltage on a capacitor (representing the membrane potential of a neuronal soma), which causes the resistive memory device to switch to a low resistance state. The resulting current increase was converted into either a shaped analog spike or a digital event by external circuitry. The randomness associated with the resistive memory switching leads to a stochastically firing neuron; a probabilistic winner-take-all network (as described in [204]) was simulated. A similar circuit, but using a $Cu/Ti/Al_2O_3$-based CBRAM device, was proposed by Jang et al. [205].

Resistive memory devices have also found applications in emulations of axonal behavior. The neuristor was first proposed as an electronic analogue to the Hodgkin–Huxley axon [206,207], but early implementations were not scalable. Pickett et al. demonstrated a neuristor built using two nanoscale Mott memristors – dynamical devices that exhibit transient memory and negative differential resistance due to an insulating-to-conducting phase transition driven by Joule heating [191]. By exploiting the functional similarity between the dynamical resistance behavior of Mott memristors and Hodgkin–Huxley Na^+ and K^+ ion channels, a neuristor comprising two NbO_2 memristors was shown to exhibit the important neuronal functions of all-or-nothing spiking with signal gain and diverse periodic spiking.

A probabilistic deep spiking neural system enabled by MTJs was proposed to transform an already fully-trained DNN into a spiking neural network (SNN) for forward-inference [186]. DNN inputs were rate-encoded as Poisson spike-trains for the SNN and modulated by the synaptic weights, with the resulting post-synaptic current flowing through the heavy metal underneath the MTJ device. This current flow induced switching in the MTJ from AP to P with a probability distribution similar to a DNN sigmoid function, with 50% probability at zero input imposed by the addition of a constant bias current. A switch to the P state triggered an output spike. Stochastic micro-magnetic simulations of a large scale deep learning network architecture showed up to 97.6% test accuracy on MNIST handwritten digit recognition by the SNN forward-inference implementation (compared to 98.56% in the originally trained DNN).

Sharad et al. proposed the use of lateral spin valves and domain wall magnets (DWM) as neuron devices [182], for implementing multiply-accumulate functionality. In their first idea, two input magnets with opposite polarity, one fixed magnet and one output magnet were connected by a metal channel. Spin-torque transfer causes the output magnet to switch to the soft axis parallel to the polarity of the input magnet with the larger input, which can be detected through an MTJ. In their second idea, two magnets with fixed and opposite polarity were connected by a DWM with integrated MTJ (for detection). With one of the two magnets grounded, the other received the difference of excitatory and inhibitory currents, plus a bias current to center the DWM response. This difference current determines both the direction of current flow through the DWM and the resulting magnetic polarity, which can be sensed through the MTJ. Sharad et al. also proposed circuit integration schemes for unipolar and bipolar neurons as well as device circuit co-simulation of some common image processing applications.

Moon et al. realized a pattern-recognition neuromorphic system by combining a Mo/PCMO synapse device with a NbO_2 Insulator-Metal Transition (IMT) neuron device [208]. The Mo/PCMO device showed excellent reliability characteristics because of its high activation energy for the oxidation process. A NbO_2-based oscillator neuron device was used to implement a Hopfield network-based neuromorphic system using an 11k-bit array of Mo/PCMO synapse devices and NbO_2 IMT oscillator neurons.

5. Discussion and conclusions

We have reviewed the application of NVM arrays to parallel, distributed, neuromorphic computing. In general, NVM devices can help implement neuromorphic systems by offering compact, low-power and efficient ways to integrate a large number of incoming signals – a key feature of brain-inspired computing. In Table 1, we have listed what we feel are the key research needs in order to make significant forward progress towards NVM-based neuromorphic systems.

Table 1. In each of the areas surveyed in this manuscript, here we list key research advances that we feel will be needed to make significant forward progress towards NVM-based neuromorphic systems.

Topic	Key research needs
Spiking Neural Networks (SNNs)	• Scalable, global learning architecture that can harness local spike-dependent plasticity for network convergence, while supporting high sparsity for low-energy computation
Deep Neural Networks (DNNs)	• NVM devices exhibiting gentle, symmetric increases/decreases in conductance over a large dynamic-range, for high training accuracy on modern and future DNN problems • Massively parallel peripheral circuitry for fast evaluation/training speed
Memcomputing	• End-to-end use case vs. standard CMOS sufficiently compelling to justify implementation of new unit processes
Synapses based on . . .	
Phase-change memory	• Highly-scaled devices for low programming power with low drift
Conductive-bridging RAM	• Compelling architectures that can use high resistance-contrast but stochastic and binary synapses
Filamentary-RRAM	• Compelling architectures that can use low resistance-contrast but stochastic and binary synapses, or • Sufficient analog control over the RESET step even when filaments are narrow (for low-power and low-read current)
Non-filamentary-RRAM	• Solution to voltage–time dilemma offering low-power switching at ≪1 usec and ≫1000 secretention
Spin-based devices	• Compelling end-to-end use case supporting all-spin processing, including tight variability control
Transient effects	• Compelling role for Short-Term Plasticity within an SNN or other neuromorphic algorithm
NVM-as-neurons	• Compelling role for NVM-based neurons that can accommodate finite NVM endurance

Many researchers have studied SNN employing various types of NVM as synaptic connections. These most often implement a variant of a STDP learning rule, essentially a type of Hebbian learning in which synaptic connections between neurons that are often activated together in sequence are strengthened. Here signal accumulation is typically performed by integrate-and-fire neurons – and NVM devices can assist in both the integration and firing aspects. The temporal dynamics inherent in many NVM devices can also provide interesting neuromorphic functionality.

Although some interesting demonstrations of functionality, such as image recognition, have been performed, the absence of a robust global learning architecture to accompany the local STDP learning-rule remains an important limitation. Without such an algorithm, we are forced to characterize devices in terms of ‘pJ-per-spike’ even though we do not know how many spikes will really be needed to perform a useful computational task. In this context, further work on implementing these local learning rules in yet more different types of NVM devices is not likely to lead to forward progress.

In contrast, DNN trained with supervised learning and error backpropagation have been very successful in important real world applications such as image recognition and speech recognition. These algorithms call for the multiplication of large matrices and vectors, conventionally parallelized on GPUs. However, parallel matrix–vector multiplication can also be implemented efficiently on arrays of NVMs – where NVM conductance or conductance-pairs represent the strength of synaptic connections – in an analog fashion via Ohm's Law and Kirchhoff's Law. By avoiding the time and energy spent moving around large amounts of digitized weight data, this approach could have potential advantages in both speed and power consumption compared to GPU implementations. That said, demonstrating these advantages in actual hardware at the necessary scale – while delivering DNN performance (e.g. classification accuracies) indistinguishable to GPUs – remains a significant research challenge. For instance, modern GPUs are capable of training Convolutional Neural Networks of modest size (AlexNet, 7 layers) in approximately 550 usec per example [209], albeit while dissipating as much as 300W in just one GPU. Any NVM-based system must offer a significant speed-up, or a *very* significant power advantage, over these *existing* capabilities.

NVM crossbar arrays have high device density, potentially as high as $4F^2$ area per memory cell if an integrated selection device is used, where F is the minimum feature size. Design of CMOS neuron circuitry to effectively address these dense arrays of memory elements with the high degree of connectivity that many network architectures demand will be a challenge. Neuromorphic applications have been shown to be fairly robust to some kinds of device variability and non-ideality, but sensitive to asymmetry and nonlinearity of conductance response. An ideal NVM device should have a near-linear response over most of its conductance range, with each programming pulse changing conductance by only a small portion of the overall dynamic range. In contrast to conventional data storage, for which 'device history' is an unpleasant hindrance, an ideal NVM device for neuromorphic applications must *embrace* device history.

None of the NVM device types surveyed here completely fulfill the desired criteria. PCM can offer small and contiguous conductance increases through partial crystallization, but conductance decrease (melt-quenching of an amorphous plug) is abrupt. Adaptations using multiple conductances per synapse and periodic corrections have been developed, but are less than ideal. Resistance 'drift', or relaxation of the amorphous phase after the melt-quenching inherent in the RESET step could also be a potential problem for neuromorphic applications.

CBRAM offers large dynamic range, but the filament formation process tends to be abrupt. Filamentary RRAM suffers similar abrupt SET processes with lower dynamic range, but offer fab-friendly materials and device structures. In both cases, neuromorphic architectures that can use binary devices offering stochastic programming behavior may be useful. However, the end-to-end use case of scalable architectures that rapidly converge during learning to show best-in-the-world performance on useful problems must be fleshed out and shown to be compelling.

Non-filamentary RRAM offers truly bidirectional conductance change, but improvements are needed in linearity of this response, and in scaling devices down to small scale to see if low switching voltages, currents and energies are achievable while maintaining sufficient conductance contrast, switching speed and retention characteristics. Other device concepts, including MTJ and carbon nanotube transistor devices, have also been proposed. However, further advanced device development and/or invention of acceptable schemes to bridge the deficiencies of each particular class of NVM devices without loss of speed or power efficiency is likely to be necessary. In particular, while it is too early to be sure that Short-Term Plasticity has no role in any future neuromorphic system, researchers studying such devices should seek to make the case that such STP will in fact be not just useful, but uniquely enabling.

Finally, in addition to the use of NVM devices as synapses, several proposals have addressed the use of NVM as a neuron, serving either as soma or axon. Here unique functionality through physical effects within the devices – including temporal response characteristics – in a low-power and efficient form-factor is envisioned. Similar proposals have been made for non-neuromorphic, or memcomputing applications. In all these cases, what is really needed from the research community are end-to-end use cases in which the energy, speed, cost, or other advantages of such systems are so compelling that they readily justify the significant costs of developing and implementing brand-new semiconductor processes to implement these devices at scale in real CMOS fabs.

Though many problems remain to be solved, the application of NVM arrays to neuromorphic computing continues to be a potentially attractive solution to highly parallel, distributed processing of massive amounts of data, and thus can be expected to remain an active area of research for some time.

Acknowledgements

The authors wish to acknowledge management support from Bülent Kurdi, Winfried Wilcke, Chung Lam, Spike Narayan, Evangelos Eleftheriou, Tze-Chiang (T. C.) Chen, Wilfried Haensch, Heike Riel and Dario Gil.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

[1] Y.V. Pershin and M. Di Ventra, Neural Networks 23 (2010) p.881.
[2] M. Di Ventra and Y.V. Pershin, Nat. Phys. 9 (2013) p.200.
[3] G. Indiveri, B. Linares-Barranco, R. Legenstein, G. Deligeorgis and T. Prodromakis, Nanotechnology 24 (2013) p.384010.
[4] D. Kuzum, S. Yu and H.S.P. Wong, Nanotechnology 24 (2013) p.382001.
[5] T. Serrano-Gotarredona, T. Masquelier, T. Prodromakis, G. Indiveri and B. Linares-Barranco, Front. Neurosci. 7 (2013) p.2.
[6] A. Thomas, J. Phys. D: Appl. Phys. 46 (2013) p.093001.
[7] D. Querlioz, O. Bichler, A.F. Vincent and C. Gamrat, Proc. IEEE 103 (2015) p.1398.

[8] S. Saïghi, C.G. Mayr, T. Serrano-Gotarredona, H. Schmidt, G. Lecerf, J. Tomas, J. Grollier, S. Boyn, A.F. Vincent, D. Querlioz, S. La Barbera, F. Alibart, D. Vuillaume, O. Bichler, C. Gamrat and B. Linares-Barranco, Front. Neurosci. 9 (2015) p.51.

[9] B. Rajendran and F. Alibart, IEEE J. Emerg. Sel. Top. Circuits Syst. 6 (2016) p.198.

[10] W. Senn and S. Fusi, Phys. Rev. E 71 (2005) p.061907.

[11] K.K. Likharev, J. Nanoelectron. Optoelectron. 3 (2008) p.203.

[12] M.S. Zaveri and D. Hammerstrom, Neural Networks 24 (2011) p.291.

[13] S.B. Eryilmaz, D. Kuzum, S. Yu and H.S.P. Wong, *Device and system level design considerations for analog-non-volatile-memory based neuromorphic architectures*, in *IEDM Technical Digest*, Dec., IEEE, 2015, p.4.1.1.

[14] D. Strukov, F. Merrikh, M. Prezioso, X. Guo, B. Hoskins and K. Likharev, *Memory technologies for neural networks*, in *IEEE International Memory Workshop (IMW)*, May, IEEE, 2015, p.1.

[15] K. Likharev, A. Mayr, I. Muckra and O. Türel, Ann. New York Acad. Sci. 1006 (2003) p.146.

[16] G.W. Burr, R.S. Shenoy, K. Virwani, P. Narayanan, A. Padilla, B. Kurdi and H. Hwang, J. Vac. Sci. Technol. B 32 (2014) p.040802.

[17] P. Narayanan, G.W. Burr, R.S. Shenoy, S. Stephens, K. Virwani, A. Padilla, B.N. Kurdi and K. Gopalakrishnan, IEEE J. Electron Devices Soc. 3 (2015) p.423.

[18] O. Bichler, M. Suri, D. Querlioz, D. Vuillaume, B. DeSalvo and C. Gamrat, IEEE Trans. Electron Devices 59 (2012) p.2206.

[19] G.W. Burr, R.M. Shelby, C. di Nolfo, R. Shenoy, P. Narayanan, K. Virwani, E. Giacometti, B. Kurdi and H. Hwang, *Experimental demonstration and tolerancing of a large-scale neural network (1,65,000 synapses), using phase-change memory as the synaptic weight element*, in *IEDM Technical Digest*, IEEE, 2014, p.29.5.

[20] G.S. Snider, Nanotechnology 18 (2007) p.365202.

[21] B. Rajendran, Y. Liu, J.S. Seo, K. Gopalakrishnan, L. Chang, D.J. Friedman and M.B. Ritter, IEEE Trans. Electron Devices 60 (2013) p.246.

[22] L. Deng, D. Wang, Z. Zhang, P. Tang, G. Li and J. Pei, Phys. Lett. A 380 (2015) p.903.

[23] S. Yu, P.Y. Chen, Y. Cao, L. Xia, Y. Wang and H. Wu, *Scaling-up resistive synaptic arrays for neuro-inspired architecture: challenges and prospect (invited)*, in *IEDM Technical Digest*, IEEE, 2015, p 17.3.

[24] H. Markram, J. Lübke, M. Frotscher and B. Sakmann, Science 275 (1997) p.213.

[25] H. Markram, W. Gerstner and P.J. Sjöström, Front. Synaptic Neurosci. 3 (2011) p.00004.

[26] G.Q. Bi and M.M. Poo, J. Neurosci. 18 (1998) p.10464.

[27] A. Morrison, M. Diesmann and W. Gerstner, Biol. Cybern. 98 (2008) p.459.

[28] D.O. Hebb, *The Organization of Behavior*, Wiley, New York, 1949.

[29] S. Löwel and W. Singer, Am. Assoc. Adv. Sci. 255 (1992) p.209.

[30] A. Grüning and S.M. Bohte, *Spiking Neural Networks: Principles and Challenges, in European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, Elsevier, 2014.

[31] J. Vreeken, *Spiking neural networks, an introduction*, in *Utrecht University Technical Report UU-CS-2003-008*, Utrecht University, Utrecht, 2002.

[32] C. Clopath, L. Bsing, E. Vasilaki and W. Gerstner, Nat. Neurosci. 13 (2010) p.344.

[33] S. Song, K.D. Miller and L.F. Abbott, Nat. Neurosci. 3 (2000) p.919.

[34] J.M. Brader, W. Senn and S. Fusi, Neural Comput. 19 (2007) p.2881.

[35] Y. LeCun, L. Bottou, Y. Bengio and P. Haffner, Proc. IEEE 86 (1998) p.2278.

[36] P.U. Diehl and M. Cook, Front. Comput. Neurosci. 9 (2015) p.99.

[37] T. Masquelier and S.J. Thorpe, PLoS Comput. Biol. 3 (2007) p.247.

[38] T. Masquelier, R. Guyonneau and S.J. Thorpe, PLoS One 3 (2008) p.e1377.

[39] D. Neil, M. Pfeiffer and S.C. Liu, Proceedings of the 31st annual ACM Symposium on Applied Computing, (2016) p.293.

[40] P. O'Connor, D. Neil, S.C. Liu, T. Delbruck and M. Pfeiffer, Front. Neurosci. 7 (2013) p.1.

[41] E.O. Neftci, B.U. Pedroni, S. Joshi, M. Al-Shedivat and G. Cauwenberghs, Front. Neurosci. 10 (2016) p.241.

[42] O. Bichler, D. Querlioz, S.J. Thorpe, J.P. Bourgoin and C. Gamrat, *Unsupervised features extraction from asynchronous silicon retina through spike-timing-dependent plasticity*, in *Proceedings of International Joint Conference on Neural Networks*, IEEE, 2011, p.859.

[43] M. Beyeler, N.D. Dutt and J.L. Krichmar, Neural Networks 48 (2013) p.109.

[44] D. Querlioz, W.S. Zhao, P. Dollfus, J.O. Klein, O. Bichler and C. Gamrat, *Bioinspired networks with nanoscale memristive devices that combine the unsupervised and supervised learning approaches*, in *IEEE/ACM International Symposium on Nanoscale Architectures (NANOARCH)*, IEEE, 2011 (2012) p.203.

[45] D. Querlioz, O. Bichler and C. Gamrat, *Simulation of a memristor-based spiking neural network immune to device variations*, in *Proceedings of International Joint Conference on Neural Networks*, IEEE, 2011, p.1775.

[46] L. Chen, C. Li, T. Huang, X. He, H. Li and Y. Chen, *STDP learning rule based on memristor with STDP property*, in *International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2014, p.1.

[47] D. Querlioz, O. Bichler, P. Dollfus and C. Gamrat, IEEE Trans. Nanotechnol. 12 (2013) p.288.

[48] O. Bichler, D. Roclin, C. Gamrat and D. Querlioz, *Design exploration methodology for memristor-based spiking neuromorphic architectures with the xnet event-driven simulator*, in *IEEE/ACM International Symposium Nanoscale Architectures (NANOARCH 2013)*, IEEE, 2013, p.7.

[49] S. Kim, M. Ishii, S. Lewis, T. Perri, M. BrightSky, W. Kim, R. Jordan, G. Burr, N. Sosa, A. Ray, J.P. Han, C. Miller, K. Hosokawa and C. Lam, *NVM neuromorphic core with 64k-cell (256-by-256) phase change memory synaptic array with on-chip neuron circuits for continuous in-situ learning*, in *IEDM Technical Digest*, IEEE, 2015, p.17.1.

[50] Z. Wang, S. Ambrogio, S. Balatti and D. Ielmini, Front. Neurosci. 8 (2016) p.438.

[51] G. Indiveri, E. Chicca and R. Douglas, IEEE Trans. Neural Networks Learn. Syst. 17 (2006) p.211.

[52] G. Indiveri, F. Corradi and N. Qiao, *Neuromorphic architectures for spiking deep neural networks (invited)*, in *IEDM Technical Digest*, IEEE, 2015, p.4.2.

[53] G. Indiveri and S.C. Liu, Proc. IEEE 103 (2015) p.1379.

[54] N. Qiao, H. Mostafa, F. Corradi, M. Osswald, F. Stefanini, D. Sumislawska and G. Indiveri, Front. Neurosci. 9 (2015) p.141.

[55] K. Meier, *A mixed-signal universal neuromorphic computing system (invited)*, in *IEDM Technical Digest*, IEEE, 2015, p.4.6.

[56] J. Seo, B. Brezzo, Y. Liu, B.D. Parker, S.K. Esser, R.K. Montoye, B. Rajendran, J.A. Tierno, L. Chang, D.S. Modha and D.J. Friedman, *A 45nm CMOS neuromorphic chip with a scalable architecture for learning in networks of spiking neurons*, in *IEEE Custom Integrated Circuits Conference*, Sep., IEEE, 2011, p.1.

[57] F. Akopyan, J. Sawada, A. Cassidy, R. Alvarez-Icaza, J. Arthur, P. Merolla, N. Imam, Y. Nakamura, P. Datta, G.J. Nam, B. Taba, M. Beakes, B. Brezzo, J.B. Kuang, R. Manohar, W.P. Risk, B. Jackson and D.S. Modha, IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. 34 (2015) p.1537.

[58] J.K. Kim, P. Knag, T. Chen and Z. Zhang, *A 6.67mW sparse coding aSIC enabling on-chip learning and inference*, in *Symposium on VLSI Circuits*, IEEE, 2014, p.1.

[59] G. Orchard, X. Lagorce, C. Posch, S.B. Furber, R. Benosman and F. Galluppi, *Real-time event-driven spiking neural network object recognition on the spiNNaker platform*, in *IEEE International Symposium on Circuits and Systems (ISCAS)*, IEEE, 2015, 2015, p.2413.

[60] J. Bill, K. Schuch, D. Brüderle, J. Schemmel, W. Maass and K. Meier, Front. Comput. Neurosci. 4 (2010) p.129.

[61] T. Pfeil, T.C. Potjans, S. Schrader, W. Potjans, J. Schemmel, M. Diesmann and K. Meier, Front. Neurosci. 6 (2012) p.1.

[62] P.A. Merolla, J.V. Arthur, R. Alvarez-Icaza, A.S. Cassidy, J. Sawada, F. Akopyan, B.L. Jackson, N. Imam, C. Guo, Y. Nakamura, B. Brezzo, I. Vo, S.K. Esser, R. Appuswamy, B. Taba, A. Amir, M.D. Flickner, W.P. Risk, R. Manohar and D.S. Modha, Science 345 (2014) p.668.

[63] B.V. Benjamin, P. Gao, E. McQuinn, S. Choudhary, A.R. Chandrasekaran, J.M. Bussat, R. Alvarez-Icaza, J.V. Arthur, P.A. Merolla and K. Boahen, Proc. IEEE 102 (2014) p.699.

[64] D. Rumelhart, G.E. Hinton and J.L. McClelland, *A general framework for parallel distributed processing*, in *Parallel Distributed Processing*, Chap. 2, MIT Press, 1986, p.45.

[65] Y. LeCun, Y. Bengio and G. Hinton, Nature 521 (2015) p.436.

[66] T. Shibata and T. Ohmi, *Neural Microelectronic*, *IEDM Technical Digest*, Dec., IEEE, 1997, p.337.

[67] T. Morie and Y. Amemiya, IEEE J. Solid-State Circuits 29 (1994) p.1086.

[68] G.W. Burr, R.M. Shelby, S. Sidler, C. di Nolfo, J. Jang, I. Boybat, R.S. Shenoy, P. Narayanan, K. Virwani, E.U. Giacometti, B. Kurdi and H. Hwang, IEEE Trans. Electron Devices 62 (2015) p.3498.

[69] G.W. Burr, P. Narayanan, R.M. Shelby, S. Sidler, I. Boybat, C. di Nolfo and Y. Leblebici, *Large-scale neural networks implemented with nonvolatile memory as the synaptic weight element: comparative performance analysis (accuracy, speed, and power)*, in *IEDM Technical Digest*, IEEE, 2015, p.4.4.

[70] T. Gokmen and Y. Vlasov, *Acceleration of deep neural network training with resistive cross-point devices*, 2016. Available at arxiv.org/abs/1603.07341

[71] F. Alibart, E. Zamanidoost and D.B. Strukov, Nat. Commun. 4 (2013) p.2072.

[72] P.Y. Chen, B. Lin, I.T. Wang, T.H. Hou, J. Ye, S. Vrudhula, J.S. Seo, Y. Cao and S. Yu, Mitigating effects of non-ideal synaptic device characteristics for on-chip learning, in *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design*, ICCAD '15, IEEE Press, Austin, 2015, p.194.

[73] J.W. Jang, S. Park, G.W. Burr, H. Hwang and Y.H. Jeong, IEEE Electron Device Lett. 36 (2015) p.457.

[74] F. Alibart, L. Gao, B.D. Hoskins and D.B. Strukov, Nanotechnology 23 (2012) p.075201.

[75] L. Gao, P.Y. Chen and S. Yu, IEEE Electron Device Lett. 36 (2015) p.1157.

[76] Z. Xu, A. Mohanty, P. Y. Chen, D. Kadetotad, B. Lin, J. Ye, S. Vrudhula, S. Yu, J.S. Seo and Y. Cao, *Parallel programming of resistive cross-point array for synaptic plasticity*, in *BICA 2014. 5th Annual International Conference on Biologically Inspired Cognitive Architectures*, Vol. 41, BICA Society, 2014, p.126.

[77] L. Gao, I.T. Wang, P.Y. Chen, S. Vrudhula, J.S. Seo, Y. Cao, T.H. Hou and S. Yu, Nanotechnology 26 (2015) p.455204.

[78] M.V. Nair, M.V. Nair and P. Dudek, *Gradient-descent-based learning in memristive crossbar arrays*, in *International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2015, p.1.

[79] P.M. Sheridan, C. Du and W.D. Lu, IEEE Trans. Neural Networks Learn. Syst. PP. (2016), p.1.

[80] S. Choi, P. Sheridan and W.D. Lu, Sci. Rep. 5 (2015) p.10492.

[81] M.N. Bojnordi and E. Ipek, *Memristive Boltzmann machine: a hardware accelerator for combinatorial optimization and deep learning*, in *IEEE International Symposium on High Performance Computer Architecture (HPCA)*, IEEE, 2016, p.1.

[82] S. Sidler, I. Boybat, R.M. Shelby, P. Narayanan, J. Jang, A. Fumarola, K. Moon, Y. Leblebici, H. Hwang and G.W. Burr, *Large-scale neural networks implemented with non-volatile memory as the synaptic weight element: impact of conductance response*, in *Proceeding of the European Solid-State Device Research Conference (ESSDERC)*, IEEE, 2016, p.1.

[83] D. Chabi, D. Querlioz, W.S. Zhao and J.O. Klein, ACM J. Emerg. Technol. Comput. Syst. 10 (2014) p.5.

[84] C. Gamrat, O. Bichler and D. Roclin, *Memristive based device arrays combined with spike based coding can enable efficient implementations of embedded neuromorphic circuits (invited)*, in *IEDM Technical Digest*, IEEE, 2015, p.4.5.

[85] D. Garbin, O. Bichler, E. Vianello, Q. Rafhay, C. Gamrat, L. Perniola, G. Ghibaudo and B. DeSalvo, *Variability-tolerant convolutional neural network for pattern recognition applications based on OxRAM synapses*, in *IEEE International Electron Devices Meeting*, Dec., IEEE, 2014, p.28.4.1.

[86] D. Garbin, E. Vianello, O. Bichler, Q. Rafhay, C. Gamrat, G. Ghibaudo, B. DeSalvo and L. Perniola, IEEE Trans. Electron Devices 62 (2015) p.2494.

[87] M. Prezioso, F. Merrikh-Bayat, B.D. Hoskins, G.C. Adam, K.K. Likharev and D.B. Strukov, Nature 521 (2015) p.61.

[88] M. Prezioso, I. Kataeva, F. Merrikh-Bayat, B. Hoskins, G. Adam, T. Sota, K. Likharev and D. Strukov, *Modeling and implementation of firing-rate neuromorphic-network classifiers with bilayer Pt/Al_2O_3/TiO_{2-x}/Pt Memristors*, in *IEDM Technical Digest*, Dec., IEEE, 2015, p.17.4.1.

[89] F. Merrikh, B. Hoskins and D.B. Strukov, Appl. Phys. A. 118 (2015), p.779.

[90] I. Kataeva, F. Merrikh-Bayat, E. Zamanidoost and D. Strukov, Efficient training algorithms for neural networks based on memristive crossbar circuits, in *International Joint Conference on Neural Networks (IJCNN)*, Jul., IEEE, 2015, p. 1.

[91] J. Borghetti, G.S. Snider, P.J. Kuekes, J.J. Yang, D.R. Stewart and R.S. Williams, Nature 464 (2010) p.873.

[92] S. Kvatinsky, G. Satat, N. Wald, E.G. Friedman, A. Kolodny and U.C. Weiser, IEEE Trans. Very Large Scale Integr. Syst. 22 (2014) p.2054.

[93] B. Chen, F. Cai, J. Zhou, W. Ma, P. Sheridan and W.D. Lu, *Efficient in-memory computing architecture based on crossbar arrays*, in *IEDM Technical Digest*, IEEE, 2015, p.17.5.1.

[94] S. Kvatinsky, D. Belousov, S. Liman, G. Satat, N. Wald, E.G. Friedman, A. Kolodny and U.C. Weiser, IEEE Trans. Circuits Syst. Express Briefs 61 (2014) p.895.

[95] N. Talati, S. Gupta, P. Mane and S. Kvatinsky, IEEE Trans. Nanotechnol. 15 (2016) p.635.

[96] A.V. Ievlev, S. Jesse, A.N. Morozovska, E. Strelcov, E.A. Eliseev, Y.V. Pershin, A. Kumar, V.Y. Shur and S.V. Kalinin, Nat. Phys. 10 (2014) p.59.

[97] M. Cassinerio, N. Ciocchini and D. Ielmini, Adv. Mater. 25 (2013) p.5975.

[98] D. Loke, J.M. Skelton, W.J. Wang, T.H. Lee, R. Zhao, T.C. Chong and S.R. Elliott, Proc. Nat. Acad. Sci. 111 (2014) p.13272.

[99] C.D. Wright, Y. Liu, K.I. Kohary, M.M. Aziz and R.J. Hicken, Adv. Mater. 23 (2011) p.3408.

[100] C.D. Wright, P. Hosseini and J.A.V. Diosdado, Adv. Funct. Mater. 23 (2013) p.2248.

[101] P. Hosseini, A. Sebastian, N. Papandreou, C.D. Wright and H. Bhaskaran, Electron Device Lett. 36 (2015) p.975.

[102] H. Xu, Y. Xia, K. Yin, J. Lu, Q. Yin, J. Yin, L. Sun and Z. Liu, Sci. Rep. 3 (2013) p.1230.

[103] R. Pappu, B. Recht, J. Taylor and N. Gershenfeld, Science 297 (2002) p.2026.

[104] A. Alaghi and J.P. Hayes, ACM Trans. Embedded Comput. Syst. 12 (2013) p.92.

[105] A. Fukushima, T. Seki, K. Yakushiji, H. Kubota, H. Imamura, S. Yuasa and K. Ando, Applied Physics Express 7 (2014) p.083001.

[106] P. Knag, W. Lu and Z. Zhang, IEEE Transactions on Nanotechnology 13 (2014) p.283.

[107] S. Balatti, S. Ambrogio, Z. Wang and D. Ielmini, IEEE Journal on Emerging and Selected Topics in Circuits and Systems 5 (2015).

[108] M. Le, T. Tuma, F. Zipoli, A. Sebastian and E. Eleftheriou, *Inherent stochasticity in phase-change memory devices*, in *Proceeding of the European Solid-State Device Research Conference (ESSDERC)*, IEEE, 2016.

[109] S. Balatti, S. Ambrogio, R. Carboni, V. Milo, Z. Wang, A. Calderoni, N. Ramaswamy and D. Ielmini, IEEE Transactions on Electron Devices 63 (2016) p.2029.

[110] Y.V. Pershin and M. Di Ventra, Physical Review E 84 (2011) p.046703.

[111] F.L. Traversa and M. Di Ventra, IEEE Transactions on Neural Networks and Learning Systems 26 (2015) p.2702.

[112] M. Di Ventra and Y.V. Pershin, Scientific American 312 (2015) p.56.

[113] I.L. Markov, *A review of "mem-computing np-complete problems in polynomial time using polynomial resources"*, 2014, arXiv preprint arXiv:1412.0650

[114] S. Raoux, G.W. Burr, M.J. Breitwisch, C.T. Rettner, Y.C. Chen, R.M. Shelby, M. Salinga, D. Krebs, S.H. Chen, H.L. Lung and C.H. Lam, IBM Journal of Research and Development 52 (2008) p.465.

[115] G.W. Burr, M.J. Breitwisch, M. Franceschini, D. Garetto, K. Gopalakrishnan, B. Jackson, B. Kurdi, C. Lam, L.A. Lastras, A. Padilla, B. Rajendran, S. Raoux and R. Shenoy, J. Vac. Sci. Technol. B 28 (2010) p.223.

[116] G.W. Burr, M.J. Brightsky, A. Sebastian, H.Y. Cheng, J.Y. Wu, S. Kim, N.E. Sosa, N. Papandreou, H.L. Lung, H. Pozidis, E. Eleftheriou and C.H. Lam, IEEE J. Emerg. Sel. Top. Circuits Syst. 6 (2016) p.146.

[117] M.N. Kozicki, M. Park and M. Mitkova, IEEE Trans. Nanotechnol. 4 (2005) p.331.

[118] I. Valov, R. Waser, J.R. Jameson and M.N. Kozicki, Nanotechnology 22 (2011) p.254003.

[119] H.S.P. Wong, H.Y. Lee, S.M. Yu, Y.S. Chen, Y. Wu, P.S. Chen, B. Lee, F.T. Chen and M.J. Tsai, Proc. IEEE 100 (2012) p.1951.

[120] S.B. Eryilmaz, D. Kazum, S. Yu and H. Wong, *Device and system level design considerations for analog-non-volatile-memory based neuromorphic architectures (invited)*, in *IEDM Technical Digest*, IEEE, 2015, p.4.1.

[121] M. Suri, O. Bichler, D. Querlioz, O. Cueto, L. Perniola, V. Sousa, D. Vuillaume, C. Gamrat and B. DeSalvo, *Phase change memory as synapse for ultra-Dense neuromorphic systems: application to complex visual pattern extraction*, in *IEDM Technical Digest*, IEEE, 2011, p.4.4.

[122] M. Suri, O. Bichler, Q. Hubert, L. Perniola, V. Sousa, C. Jahan, D. Vuillaume, C. Gamrat and B. Desalvo, *Interface engineering of PCM for improved synaptic performance in neuromorphic systems*, in *IMW*, IEEE, 2012, p.155.

[123] M. Suri, O. Bichler, Q. Hubert, L. Perniola, V. Sousa, C. Jahan, D. Vuillaume, C. Gamrat and B. Desalvo, Solid-State Electron. 79 (2013) p.227.

[124] M. Suri, O. Bichler, D. Querlioz, B. Traore, O. Cueto, L. Perniola, V. Sousa, D. Vuillaume, C. Gamrat and B. Desalvo, J. Appl. Phys. 112 (2012) p.054904.

[125] D. Kuzum, R.D. Jeyasingh and H.S. Wong, *Energy efficient programming of nanoelectronic synaptic devices for large-scale implementation of associative and temporal sequence learning*, in *IEDM Technical Digest*, IEEE, 2011, p.30.3.

[126] D. Kuzum, R.G.D. Jeyasingh, S.M. Yu and H.S.P. Wong, IEEE Trans. Electron Devices 59 (2012) p.3489.

[127] D. Kuzum, R.G.D. Jeyasingh, B. Lee and H.S.P. Wong, Nano Lett. 12 (2012) p.2179.

[128] J.M. Skelton, D. Loke, T. Lee and S.R. Elliott, ACS Appl. Mater. Interfaces 7 (2015) p.14223.

[129] S.B. Eryilmaz, D. Kuzum, G.D. Jeyasingh, S.B. Kim, M. BrightSky, C. Lam and H.S.P. Wong, *Experimental demonstration of array-level learning with phase change synaptic devices*, in *IEDM Technical Digest*, IEEE, 2013, p.25.5.

[130] S. Ambrogio, N. Ciocchini, M. Laudato, V. Milo, A. Pirovano, P. Fantini and D. Ielmini, Front. Neurosci. 10 (2016) p.56.

[131] Y. Li, Y.P. Zhong, L. Xu, J.J. Zhang, X.H. Xu, H.J. Sun and X.S. Miao, Sci. Rep. 3 (2013) p.1619.

[132] Y. Zhong, Y. Li, L. Xu and X. Miao, Phys. Status Solidi-Rapid Res. Lett. 9 (2015) p.414.

[133] B.L. Jackson, B. Rajendran, G.S. Corrado, M. Breitwisch, G.W. Burr, R. Cheek, K. Gopalakrishnan, S. Raoux, C.T. Rettner, A. Padilla, A.G. Schrott, R.S. Shenoy, B.N. Kurdi, C.H. Lam and D.S. Modha, ACM J. Emerg. Technol. Comput. Syst. 9 (2013) p.12.

[134] T. Ohno, T. Hasegawa, A. Nayak, T. Tsuruoka, J.K. Gimzewski and M. Aono, Appl. Phys. Lett. 99 (2011) p.203108.

[135] T. Ohno, T. Hasegawa, T. Tsuruoka, K. Terabe, J.K. Gimzewski and M. Aono, Nat. Mater. 10 (2011) p.591.

[136] S. Yu and S.P. Wong, *Modeling the switching dynamics of programmable-metallization-cell (PMC) memory and its application as synapse device for a neuromorphic computation system*, in *IEDM Technical Digest*, IEEE, 2010, p.22.1.

[137] S.H. Jo, T. Chang, I. Ebong, B.B. Bhadviya, P. Mazumder and W. Lu, Nano Lett. 10 (2010) p.1297.

[138] M. Suri, O. Bichler, D. Querlioz, G. Palma, E. Vianello, D. Vuillaume, C. Gamrat and B. DeSalvo, *CBRAM devices as binary synapses for low-power stochastic neuromorphic systems: auditory (cochlea) and visual (retina) cognitive processing applications*, in *IEDM Technical Digest*, IEEE, 2012, p.10.3.

[139] M. Ziegler, R. Soni, T. Patelczyk, M. Ignatov, T. Bartsch, P. Meuffels and H. Kohlstedt, Adv. Funct. Mater. 22 (2012) p.2744.

[140] Y. Yang, B. Chen and W.D. Lu, Adv. Mater. 27 (2015) p.7720.

[141] H.O. Sillin, R. Aguilera, H.H. Shieh, A.V. Avizienis, M. Aono, A.Z. Stieg and J.K. Gimzewski, Nanotechnology 24 (2013) p.384004.

[142] H. Choi, H. Jung, J. Lee, J. Yoon, J. Park, D.J. Seong, W. Lee, M. Hasan, G.Y. Jung and H. Hwang, Nanotechnology 20 (2009) p.345201.

[143] S. Yu, Y. Wu, R. Jeyasingh, D.G. Kuzum and H.S.P. Wong, IEEE Trans. Electron Devices 58 (2011) p.2729.

[144] Y. Wu, S. Yu, H.S.P. Wong, Y.S. Chen, H.Y. Lee, S.M. Wang, P.Y. Gu, F. Chen and M.J. Tsai, *AlOx-based resistive switching device with gradual resistance modulation for neuromorphic device application*, in *IMW*, IEEE, 2012, p.111.

[145] M. Itoh and L.O. Chua, Bifurcation Chaos 19 (2009) p.3605.

[146] S.P. Adhikari, C. Yang, H. Kim and L.O. Chua, IEEE Trans. Neural Networks Learn. Syst. 23 (2012) p.1426.

[147] H. Kim, M.P. Sah, C. Yang, T. Roska and L.O. Chua, Proc. IEEE 100 (2012) p.2061.

[148] S.M. Yu, B. Gao, Z. Fang, H.Y. Yu, J.F. Kang and H.S.P. Wong, Adv. Mater. 25 (2013) p.1774.

[149] S. Yu, B. Gao, Z. Fang, H.Y. Yu, J.F. Fang and H.S.P. Wong, *A neuromorphic visual system using RRAM synaptic devices with Sub-pJ energy and tolerance to variability: experimental characterization and large-scale modeling*, in *IEDM Technical Digest*, IEEE, 2012, p.10.4.

[150] S. Yu, B. Gao, Z. Fang, H. Yu, J. Kang and H.S.P. Wong, Front. Neurosci. 7 (2013) p.186.

[151] Y.J. Jeong, S. Kim and W.D. Lu, Appl. Phys. Lett. 107 (2015) p.173105.

[152] B. Sarkar, B. Lee and V. Misra, Semicond. Sci. Technol. 30 (2015) p.105014.

[153] C. Wang, W. He, Y. Tong and R. Zhao, Sci. Rep. 6 (2016) p.22970.

[154] S. Ambrogio, S. Balatti, F. Nardi, S. Facchinetti and D. Ielmini, Nanotechnology 24 (2013) p.384012.

[155] K.H. Kim, S. Gaba, D. Wheeler, J.M. Cruz-Albrecht, T. Hussain, N. Srinivasa and W. Lu, Nano Lett. 12 (2012) p.389.

[156] J. Bill and R. Legenstein, Front. Neurosci. 8 (2014) p.412.

[157] G. Piccolboni, G. Molas, J.M. Portal, R. Coquand, M. Bocquet, D. Garbin, E. Vianello, C. Carabasse, V. Delaye, C. Pellissier, T. Magis, C. Cagli, M. Gely, O. Cueto, D. Deleruyelle, G. Ghibaudo, B.De Salvo and L. Perniola, *Investigation of the potentialities of vertical resistive RAM (VRRAM) for neuromorphic applications*, in *IEDM Technical Digest*, IEEE, 2015, p.17.2.

[158] M. Prezioso, F. Merrikh, B. Hoskins, K. Likharev and D. Strukov, *Self-adaptive spike-time-dependent plasticity of metal-oxide memristors*, 2015, arXiv preprint arXiv:1505.05549

[159] J.M. Cruz-Albrecht, T. Derosier and N. Srinivasa, Nanotechnology 24 (2013) p.384011.

[160] L. Deng, G. Li, N. Deng, D. Wang, Z. Zhang, W. He, H. Li, J. Pei and L. Shi, Sci. Rep. 5 (2015).

[161] R. Waser, R. Dittmann, G. Staikov and K. Szot, Adv. Mater. 21 (2009) p.2632.

[162] D.B. Strukov, G.S. Snider, D.R. Stewart and R.S. Williams, Nature 453 (2008) p.80.

[163] K. Seo, I. Kim, S. Jung, M. Jo, S. Park, J. Park, J. Shin, K.P. B, J. Kong, K. Lee, B. Lee and H. Hwang, Nanotechnology 22 (2011) p.254023.

[164] T. Chang, S.H. Jo, K.H. Kim, P. Sheridan, S. Gaba and W. Lu, Appl. Phys. A 102 (2011) p.857.

[165] I.T. Wang, Y.C. Lin, Y.F. Wang, C.W. Hsu and T.H. Hou, *3D synaptic architecture with ultralow sub-10 fJ energy per spike for neuromorphic computation*, in *IEDM Technical Digest*, IEEE, 2014, p.28.5.

[166] Y.F. Wang, Y.C. Lin, I.T. Wang, T.P. Lin and T.H. Hou, Sci. Rep. 5 (2015) p.10150.

[167] H. Lim, I. Kim, J.S. Kim, C.S. Hwang and D.S. Jeong, Nanotechnology 24 (2013) p.384005.

[168] R. Yang, K. Terabe, Y. Yao, T. Tsuruoka, T. Hasegawa, J.K. Gimzewski and M. Aono, Nanotechnology 24 (2013) p.384003.

[169] S. Park, H. Kim, M. Choo, J. Noh, A. Sheri, S. Jung, K. Seo, J. Park, S. Kim, W. Lee, J. Shin, D. Lee, G. Choi, J. Woo, E. Cha, J. Jang, C. Park, M. Jeon, B. Lee and H. Hwang, *RRAM-based synapse for neuromorphic system with pattern recognition function*, in *IEDM Technical Digest*, IEEE, 2012, p.10.2.

[170] S. Park, J. Noh, M.L. Choo, A.M. Sheri, M. Chang, Y.B. Kim, C.J. Kim, M. Jeon, B.G. Lee, B.H. Lee and H. Hwang, Nanotechnology 24 (2013) p.384009.

[171] S. Park, A. Sheri, J. Kim, J. Noh, J.W. Jang, M. Jeon, B.G. Lee, B. Lee and H. Hwang, *Neuromorphic speech systems using advanced ReRAM-based synapse*, in *IEDM Technical Digest*, IEEE, 2013, p.25.6.

[172] S. Park, M. Siddik, J. Noh, D. Lee, K. Moon, J. Woo, B.H. Lee and H. Hwang, Semicond. Sci. Technol. 29 (2014) p.104006.

[173] A.M. Sheri, H. Hwang, M. Jeon and B.G. Lee, IEEE Trans. Ind. Electron. 61 (2014) p.2933.

[174] K. Moon, S. Park, J. Jang, D. Lee, J. Woo, E. Cha, S. Lee, J. Park, J. Song, Y. Koo and H. Hwang, Nanotechnology 25 (2014) p.495204.

[175] D. Lee, K. Moon, J. Park, S. Park and H. Hwang, Appl. Phys. Lett. 106 (2015) p.113701.

[176] S. Park, M. Chu, J. Kim, J. Noh, M. Jeon, B.H. Lee, H. Hwang, B. Lee and B.G. Lee, Sci. Rep. 5 (2015) p.10123.

[177] D. Lee, J. Park, K. Moon, J. Jang, S. Park, M. Chu, J. Kim, J. Noh, M. Jeon, B. Lee, B.G. Lee and H. Hwang, *Oxide based nanoscale analog synapse device for neural signal recognition system (invited)*, in *IEDM Technical Digest*, IEEE, 2015, p.4.7.

[178] M. Hansen, M. Ziegler, L. Kolberg, R. Soni, S. Dirkmann, T. Mussenbrock and H. Kohlstedt, Sci. Rep. 5 (2015) p.13753.

[179] F. Zahari, M. Hansen, T. Mussenbrock, M. Ziegler and H. Kohlstedt, AIMS Mater. Sci. 2 (2015) p.203.

[180] F. Alibart, S. Pleutin, D. Guerin, C. Novembre, S. Lenfant, K. Lmimouni, C. Gamrat and D. Vuillaume, Adv. Funct. Mater. 20 (2010) p.330.

[181] L.Q. Zhu, C.J. Wan, L.Q. Guo, Y. Shi and Q. Wan, Nat. Commun. 5 (2014) p.3158.

[182] M. Sharad, C. Augustine and K. Roy, *Boolean and non-Boolean computation with spin devices*, in *IEDM Technical Digest*, IEEE, 2012, p.11.6.1.

[183] V.K. Sangwan, D. Jariwala, I.S. Kim, K.S. Chen, T.J. Marks, L.J. Lauhon and M.C. Hersam, Nat. Nanotechnol. 10 (2015) p.403.

[184] A.F. Vincent, J. Larroque, N.L.N.B. Romdhane, O. Bichler, C. Gamrat, W.S. Zhao, J.O.K. Galdin-Retailleau and D. Querlioz, IEEE Trans. Biomed. Circuits Syst. 9 (2015) p.166.

[185] A. Sengupta, Z. Al Azim, X. Fong and K. Roy, Appl. Phys. Lett. 106 (2015) p.093704.

[186] A. Sengupta, M. Parsa, B. Han and K. Roy, *Probabilistic deep spiking neural systems enabled by magnetic tunnel junction*, arXiv:1605.04494v1, 2016.

[187] K. Gacem, J.M. Retrouvey, D. Chabi, A. Filoramo, W. Zhao, J.O. Klein and V. Derycke, Nanotechnology 24 (2013) p.384013.

[188] A.M. Shen, C.L. Chen, K. Kim, B. Cho, A. Tudor and Y. Chen, ACS Nano 7 (2013) p.6117.

[189] S. Kim, J. Yoon, H.D. Kim and S.J. Choi, ACS Appl. Mater. Interfaces 7 (2015) p.25479.

[190] Y. Kaneko, Y. Nishitani, M. Ueda and A. Tsujimura, *Neural network based on a three-terminal ferroelectric memristor to enable on-chip pattern recognition*, in *Symposium on VLSI Technology*, Japan Society of Applied Physics, 2013, p.T16.2.

[191] M.D. Pickett, G. Medeiros-Ribeiro and R.S. Williams, Nat. Mater. 12 (2013) p.114.

[192] A. Thomas, S. Niehrster, S. Fabretti, N. Shepheard, O. Kuschel, K. Küpper, J. Wollschlger, P. Krzysteczko and E. Chicca, Front. Neurosci. 9 (2015) p.241.

[193] D. Zhang, L. Zeng, K. Cao, M. Wang, S. Peng, Y. Zhang, Y. Zhang, J.O. Klein, Y. Wang and W. Zhao, IEEE Trans. Biomed. Circuits Syst. 10 (2016) p.828.

[194] T. Tuma, A. Pantazi, M. Le Gallo, A. Sebastian and E. Eleftheriou, Nat. Nanotechnol. 11 (2016) p.693.

[195] E.R. Kandel, J.H. Schwartz, T.M. Jessell, S.A. Siegelbaum and A.J. Hudspeth, *Principles of Neural Science*, Vol. 4, McGraw-hill, New York, 2000.

[196] W. Gerstner, W.M. Kistler, R. Naud and L. Paninski, *Neuronal Dynamics: From Single Neurons to Networks and Models of Cognition*, Cambridge University Press, 2014.

[197] G. Indiveri, B. Linares-Barranco, T.J. Hamilton, A. Van Schaik, R. Etienne-Cummings, T. Delbruck, S.C. Liu, P. Dudek, P. Häfliger, S. Renaud, J. Schemme, G. Cauwenberghs,

J. Arthur, K. Hynna, F. Folowosele, S. Saighi, T. Serrano-Gotarredona, J. Wijekoon, Y. Wang and K. Boahen, Front. Neurosci. 5 (2011) p.73.

[198] S.R. Ovhsinsky, *Analog Neurons and Neurosynaptic Networks*, 2006, US Patent 6,999,953.

[199] T. Tuma, M. Le Gallo, A. Sebastian and E. Eleftheriou, Electron Device Lett. 37 (2016) p.1238.

[200] A. Pantazi, S. Wozniak, T. Tuma and E. Eleftheriou, Nanotechnology 27 (2016) p.355205.

[201] B.B. Averbeck, P.E. Latham and A. Pouget, Nat. Rev. Neurosci. 7 (2006) p.358.

[202] W. Maass, Proc. IEEE 102 (2014) p.860.

[203] M. Al-Shedivat, R. Naous, G. Cauwenberghs and K.N. Salama, IEEE J. Emerg. Sel. Top. Circuits Syst. 5 (2015) p.242.

[204] B. Nessler, M. Pfeiffer, L. Buesing and W. Maass, PLoS Comput. Biol. 9 (2013) p.e1003037.

[205] J.W. Jang, B. Attarimashalkoubeh, A. Prakash, H. Hwang and Y.H. Jeong, IEEE Trans. Electron Devices 63 (2016) p.1.

[206] A.L. Hodgkin and A.F. Huxley, J. Physiol. 117 (1952) p.500.

[207] H.D. Crane, Proc. IRE 50 (1962) p.2048.

[208] K. Moon, E. Cha, J. Park, S. Gi, M. Chu, K. Baek, B. Lee, S. Oh and H. Hwang, *High density neuromorphic system with Mo/$Pr_xCa_{1-x}MnO_3$ synapse and NbO_2 IMT oscillator neuron*, in *IEDM Technical Digest*, IEEE, 2015, p.17.6.

[209] S. Chintala, *Convnet benchmarks*, 2016, Available at https://github.com/soumith/convnet-benchmarks