TOPICAL REVIEW

Device and materials requirements for neuromorphic computing

IOP Publishing Journal of Physics D: Applied Physics
J. Phys. D: Appl. Phys. **52** (2019) 113001 (24pp) https://doi.org/10.1088/1361-6463/aaf784

Topical Review

Device and materials requirements for neuromorphic computing

Raisul Islam[1], Haitong Li[1], Pai-Yu Chen[2], Weier Wan[1], Hong-Yu Chen[3], Bin Gao[4], Huaqiang Wu[4], Shimeng Yu[5], Krishna Saraswat[1] and H-S Philip Wong[1]

[1] Department of Electrical Engineering, Stanford University, 420 Via Palou Mall, Stanford, CA 94305, United States of America
[2] School of Electrical, Computer and Energy Engineering, Arizona State University, Tempe, AZ 85287, United States of America
[3] GigaDevice Semiconductor Inc., A12, USTB Techart Plaza, Xueyuan Road 30, Haidian District, Beijing, People's Republic of China
[4] Institute of Microelectronics, Tsinghua University, Beijing 100084, People's Republic of China
[5] School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332, United States of America

E-mail: raisul@stanford.edu and hspwong@stanford.edu

Received 5 August 2018, revised 12 October 2018
Accepted for publication 10 December 2018
Published 18 January 2019

Abstract

Energy efficient hardware implementation of artificial neural network is challenging due the 'memory-wall' bottleneck. Neuromorphic computing promises to address this challenge by eliminating data movement to and from off-chip memory devices. Emerging non-volatile memory (NVM) devices that exhibit gradual changes in resistivity are a key enabler of in-memory computing—a type of neuromorphic computing. In this paper, we present a review of some of the NVM devices (RRAM, CBRAM, PCM) commonly used in neuromorphic application. The review focuses on the trade-off between device parameters such as retention, endurance, device-to-device variation, speed and resistance levels, and the interplay with target applications. This work aims at providing guidance for finding the optimized resistive memory devices material stack suitable for neuromorphic application.

Keywords: neuromorphic computing, non volatile memory, deep neural network

(Some figures may appear in colour only in the online journal)

1. Introduction

First coined by Carver Mead in 1990 [1], the term 'neuromorphic computing' refers to a computing paradigm inspired by the cognitive functionality of human brain. In today's data-centric world, where some of the most useful computing tasks are to extract meaningful information from massive amounts of unstructured data, neuromorphic computing can provide low-energy high throughput computing. The challenge in data-centric intelligent computing with the conventional computing architecture lies in the energy and latency bottleneck of off-chip memory access (i.e. 'memory wall') which do not scale down with the scaling of the technology node [2]. To overcome this problem, new in-memory computing paradigm has been proposed [3–13] for accelerating deep neural networks (DNNs) used in data-centric computing. In-memory computing can utilize certain properties of the emerging non-volatile memory (NVM) devices such as gradual switching of resistance values with constant voltage pulse train. Besides application-oriented accelerator hardware for neural networks (NNs), neuromorphic computing may also aim at emulating brain-like learning behavior (e.g. spike timing dependent

plasticity (STDP)) in electronic systems. Conventional hardwares like CPUs and GPUs that emulate brain-like functionality is not energy efficient [14]. As an alternative, resistive memory as synaptic connection between two neurons is promising for brain-inspired computing. Such non-volatile memories with multiple levels of resistance states can be easily integrated on-chip that can be used as an analog weight storage reducing the memory access overhead. Alternatively, it can facilitate certain processing tasks to be performed in memory resulting in further reduction of memory access overhead at lower energy cost [15].

Analog-programmable NVM devices such as resistive RAM (RRAM), conductive bridging RAM (CBRAM), phase change memory (PCM), magnetic RAM (STT-MRAM) lie at the heart of such neuromorphic computing devices. A fundamental device element having resistive memory, termed as 'memristor', has been theorized by Chua *et al* [16]. Later Strukov *et al* proposed that Pt/TiO_{2-x}/Pt resistive switching devices are the physical embodiment of memristors [17]. Although these works had significant impact on the field of NVM devices for neuromorphic computing, later it was shown that the typical resistive memory devices (e.g. RRAM, CBRAM, PCRAM) are not equivalent to the memristors with respect to its working principle [18] theorized by Chua *et al* [16]. NVM devices were originally developed as digital memories which could be used as on-chip memory or non-volatile data storage. However, one important capability of these devices is the multi-bit capacity where instead of two resistance levels, multiple levels can be encoded to multi-bit information. This gradual switching of the resistance levels in these devices are the key to neuromorphic applications. While extensive reviews of the emerging NVM devices for storage-class memory application exist in the literature, a comprehensive review of the devices and materials requirements and possible trade-offs for neuromorphic application is missing. Note that resistive memory devices based on organic materials fall into a different class of devices suitable for neuromorphic computing. These devices are still not matured enough to be readily available for commercial technology, yet they show promising characteristics like excellent capability of analog tuning, linearity in conductance and extremely low energy for switching. Detailed review of the state-of-the-art of such devices is presented elsewhere [19] and is out of scope for this paper. Hence, the goal of this paper is to present a review of inorganic materials based NVM devices for neuromorphic application. Two similar yet broad reviews on the relevant topic were done by Burr *et al* [20] and Yu *et al* [21]. This review is more focused towards device trade-offs for hardware artificial neural networks (ANNs) exploiting in-memory computing principles.

The paper is organized in three main sections. First, we explain the challenges of state-of-the-art hardware accelerators present in literature and show how NVM devices could be useful in such systems. Then we provide a review of the various NVM technologies that have already been demonstrated for this application. Finally, we discuss the possible device trade-offs in designing neuromorphic hardware.

2. Overview of neuromorphic computing

Neuromorphic computing can be broadly classified into two categories: (a) biology based models/algorithms which are based on studying the learning and inference process of the human brain and emulating those functionalities in hardware and (b) ANNs which are algorithms to solve machine learning problems inspired by the brain to some extent (network layers are constructed by the connections between neurons termed as synapses) but does not necessarily have a direct correlation with brain functionality. The human brain consists of neurons which are interconnected by a highly complex network of synapses. Each neuron is connected to multiple neurons through synapses. Neurons generate action potentials (spikes) that are transmitted to the other connected neurons through synapses. The communication between the neurons through spiking signals results in the modification of synaptic connection strength. This synaptic strength modulation forms the basis for learning that can be emulated in hardware (class (a)). For example, one such learning paradigm is known as STDP [22], where each neuron integrates all the incoming action potential and when the integrated signal exceeds a certain threshold, it fires a spiking pulse that contributes to the learning by changing the synaptic connection strength based on the timing of pre-syanptic and post-synaptic pulse. STDP can be a 'local' learning mechanism which is only applicable in emulating brain-like behavior. Implementation of such learning models in hardware originating directly from understanding the brain's learning mechanism has been studied and reviewed extensively by Kuzum *et al* [15]. STDPs can also be viewed as a 'global' learning mechanism that requires weight updating (via for instance error backpropagation) for neuromorphic computing applications. Our discussion in this paper will be focused on hardware acceleration of ANNs using emerging NVM devices. The ANNs are inspired from the human brain's neural connectivity, yet do not correlerate to any specific biological learning model. Figure 1 shows the development of neuromorphic computing and its categories. This paper will focus on the highlighted square of the design paradigm.

3. Hardware acceleration for NN

DNNs are a class of ANNs that features a considerable increase in the network depth to build richer representations of the input data. DNNs have been gaining great momentum for tackling large-scale, perceptional tasks such as computer vision and natural language understanding. This secion picks DNN as a case study, among many variants of ANNs, to illustrate the close interaction between the development of hardware primitives and NNs. DNNs have been benefiting from both the availability of big data (large amount of multi-media data for model training) and the large performance improvement of computing hardware in the past decade. Recent development of DNNs features an increase in both the model size (defined as the amount of static weights after training of a NN) and computational complexity (feedforward and backward) [23–27], to meet the requirements of demanding tasks such

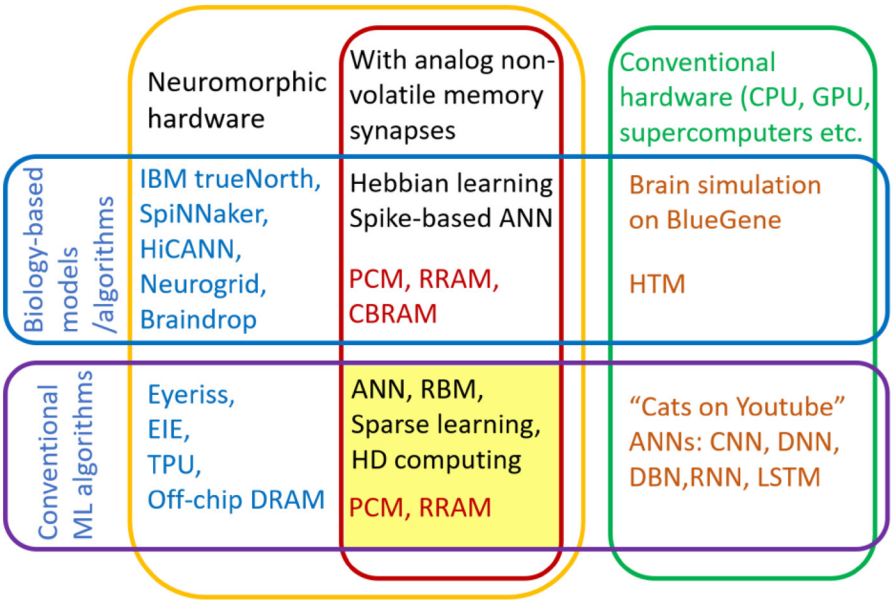

Figure 1. Neuromorphic computing paradigm. In each box, already implemented examples are given along with the NVM device technology utilized (PCM = phase change memory, RRAM = resistive RAM, CBRAM = conductive bridging RAM). The region highlighted in yellow is the topic highlighted in this paper. This figure is adapted from [80, 113, 114, 142–152].

as video processing [28]. Many recent general-purpose hardware platforms (e.g. CPU and GPU) employ special features such as vector instruction [29] and mixed-precision operations [30] to improve parallelism for DNN inference and training. However, the memory hiararchy design of these general-purpose architectures is not specifically designed to leverage the predictable dataflow and potential data reuse of DNN processing. Therefore, a large portion of memory access goes to the slower and more energy-consuming levels of memory hiararchy (e.g. one off-chip DRAM main memory consumes much more energy than local and small register files per access), limiting the compute throughput and energy efficiency of DNN processing. To reduce these expensive memory access, hardware accelerators for DNNs are designed to employ more fine-grained local memory hiararchy and more specialized dataflow design, which improves the energy efficiency and throughput while maintaining DNN's inference accuracy. Modern DNNs typically consist of convolutional (CONV) layers and fully-connected (FC) layers as trainable layers (containing weight parameters), interwoven with pre-defined non-linear activation, normalization, pooling and regularization layers that are typically not compute-intensive. However, both CONV and FC layers require intensive multiply-and-accumulate (MAC) operations during both feedforward and back-propagation computation. These MAC operations performed for millions of weight parameters in a DNN impose a stringent requirement on the efficiency of memory access. Therefore, a common target of the state-of-the-art DNN accelerator designs are two-fold: accelerating the MAC operations while minimizing the energy cost of data movement. In this section, we will first review the design and optimization methodologies of DNN accelerators. Then, the architectural implications for the use of new memory technologies in this context will be discussed.

3.1. Design and optimization methodologies

Modern DNNs are both computation-intensive and memory-intensive. As seen in some popular DNN architectures, the total number of weights is in the order of tens or hundreds of millions, while the total number of MAC operations during inference can be two to three orders of magnitude larger [23, 24, 26, 27]. For instance, a ResNet-50 network trained on the ImageNet dataset can contain over 20 M weights and require about 4G MAC operations [23]. Performing inference for a batch of 16 images using ResNet-50 on two Intel Xeon E5-2630 v3 processors takes more than 6.6 s to complete [31]. Associated with every single MAC operation during the inference phase of a DNN, there are several memory accesses on weight data, activation data, and partial-sum data before and after the computation. These memory accesses can be rather inefficient with general-purpose architectures as a substantial portion goes to relatively slow and energy-hungry, off-chip DRAM. Therefore, to address this memory wall issue and to minimize the data fetching/movement costs, the first methodology that DNN accelerators have taken is to use spatial architectures, which consist of distributed arithmetic logic units (ALUs), localized (yet capacity-limited) memories (e.g. register files, local buffers), and an on-chip network that enables direct communication between ALUs. Some of the early examples include neuFlow [32] and DianNao [33]. The former design uses local registers to store frequently-accessed weights for each MAC unit, while the latter uses scratchpad SRAMs to store weights and intermediate inputs/outputs. In addition to minimizing the energy of reading weights from memories, ShiDianNao (one of the successors of DianNao) [34] is designed to minimize memory write accesses by grouping the MAC outputs from adjacent ALUs before writing back to SRAMs. The strategies employed in [32–34] can be summarized as minimizing read and write accesses by handling,

caching, and processing reusable data in a DNN computation flow, and can be seen in several other reports of DNN accelerators as well [35, 36]. Most recently, Eyeriss combined these two strategies to further improve the data reuse, by efficiently compiling and mapping DNN parameters from DRAM into scratchpad SRAMs and local registers [7]. As NNs can be viewed as arbitrary function approximators from an algorithm perspective, weight precision reduction and network pruning may be used to compress large DNN models and yield smaller models that can better fit hardware constraints during deployment [6]. Some DNN accelerator designs have exploited this methodology by mapping compressed DNN models to reduce energy and area costs of high-precision arithmetic. As a result, these compressed models typically require less storage and compute resources on hardware. For example, EIE [6] and SCNN [11] are inference accelerators that use network pruning technique, which takes redundant weight parameters and set them to zero. EIE is designed to perform computation on the sparse representation after pruning FC layers, while SCNN focuses such sparse processing on CONV layers. Google's™ tensor processing unit (TPU) reduces the precision to 8-b integer arithmetic [37], while some other accelerators explore even less number of bits to improve throughput and energy efficiency, including ternary/binary representations [3, 38].

3.2. Architectural implications for memory technologies

Present accelerator designs put a central emphasis on the memory hierachy optimization and its interplay with on-chip computation resources. However, the 'memory bottleneck' in modern DNNs may not be fully addressed by the aforementioned acceleration architectures alone. In fact, memory access remains to be the bottleneck for many DNN inference workloads when deployed on acceleartor hardware, especially for networks mainly consisting of FC layers, such as multilayer perceptron (MLP) and long short-term memory (LSTM) [37]. Moreover, one can expect future DNNs to grow rapidly in network depth and computational complexity. As an example, DNNs with convolutional layers for image applications have grown from eight layers (AlexNet [24]) to over 100 layers (ResNet [23]) to be able to handle rich information in natural images. The state-of-the-art YOLO network [39] for real-time object detection involves computations on many small grids of a single image or video frame, which implies the growing need for more data-intensive, fine-grained multi-media processing. Thus, in the future, real-time processing of high-resolution videos would require even more hardware computing capabilities to handle parallel processing with large DNNs and the concurrent memory accesses with as little bandwidth limitation as possible. Contemporary accelerator designs still face the memory bandwidth and capacity wall, as the typical on-chip registers and SRAM buffers can only provide KB- to MB-scale data memory [7, 33, 37], which is much smaller than off-chip DRAM capacity. This has driven several accelerator works towards using alternative memory technologies. For instance, DaDianNao [5], another successor of DianNao, uses 36-MB/chip embedded DRAM (eDRAM) to provide slightly larger on-chip storage capacity compared to SRAM. However, such approach may not have good scalability, due to the added cost of eDRAM technology and limited benefits for on-chip storage capacity. For state-of-the-art node (14 nm), in high volume manufacturing ~70% array efficiency has been demonstrated for on-chip SRAM [40]. If ~80% of the chip is SRAM macros in futuristic nodes (~7 nm) where SRAM bit cell area is 0.027 μm^2 [41] a typical die (815 mm^2—NVIDIA V100 [42]) could accommodate ~2 GB of SRAM in future. 2 GB of SRAM sounds sufficient to hold most of today's DNN weights on-chip, however, in this case, the standby leakage of the SRAM array may dominate the entire chip's power comsumption, which makes it unpractical. Considering the memory wall faced by the modern DNNs, emerging memory technologies may play an important and unique role. The candidates that are being actively investigated by the device and material communities include PCM [43, 44], RRAM [45, 46], CBRAM [47], and STT-MRAM [48]. As these technologies can potentially offer up to tera-bytes of on-chip data storage with a wide range of energy-delay optimization opportunities, they may complement SRAM for more efficient DNN inference acceleration. Architecture studies, through simulations, have shown that RRAM crossbar arrays can provide MAC processing capability and on-chip data storage at the same time [8, 9]. These studies use the structual parallelism and current summation properties, but do not fully exploit the analog programmable properties of resistive-type non-volatile memories. Thus, there is an even larger design space with emerging memory technologies that can be exploited as a key compute and storage component for efficient hardware implementations of DNNs. The following sections will address this topic in detail.

3.3. NVM as analog synaptic weights in NNs

A possible application of emerging NVM devices is to serve as in-memory computing element where multi-level resistance response of an NVM can store the analog synaptic weights of a DNN on-chip. After reading these analog weights, conventional hardwares can perform the typical arithmetic operation. These schemes bring the memory closer to the computing element but the computation is not done inside the memory. In another in-memory computing scheme, a crossbar array of NVM devices can perform the MAC operation at a lower energy cost when the input vector is encoded as an analog voltage and the weight matrix is encoded as analog resistance (conductance) values stored in the memory devices. Figure 2 shows the typical mathematical abstraction of a single layer perceptron. If the input vector is encoded as an analog voltage and the weight matrix can be encoded as the conductance values in a resistive memory array (figure 2(b)), the output current represents the MAC operation. The ability of the NVM devices like RRAM, PCM, CBRAM to change its resistance values gradually as a function of the applied voltage pulse across its electrode is the key to performing analog in-memory MAC operation (figure 2). However, if the NVM device has non-linear *I–V* curve (which is typically the case in higher resistance

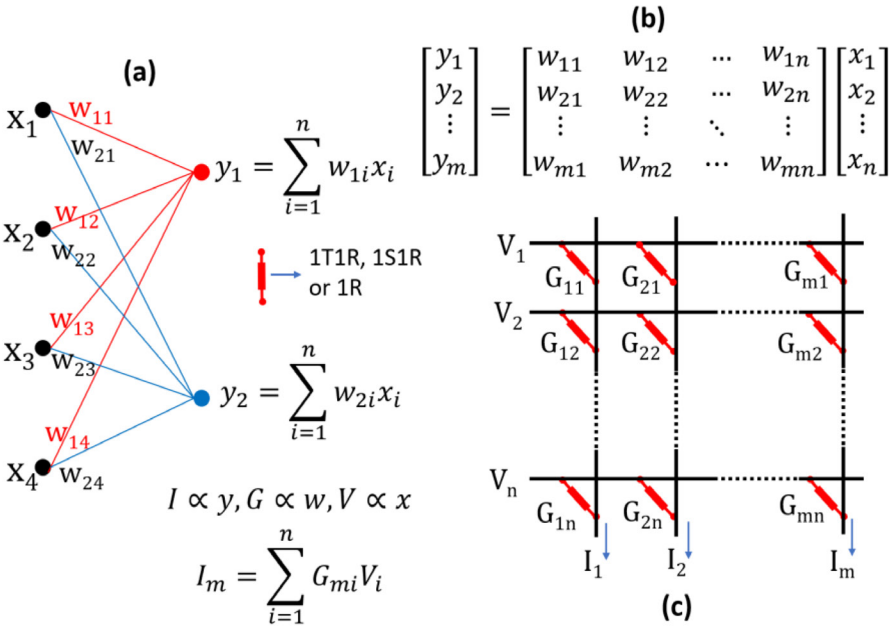

Figure 2. (a) Single layer perceptron with four inputs and two outputs. (b) General computational form for single layer ANN. (c) NVM crossbar array for realizing the matrix-vector multiplication shown in (b). Here, T = transistor, S = selector, R = resistor.

state), using analog voltage as input will cause large error due to the variable conductance with the read voltage. A solution is to use an identical pulse train as input where the pulse number represents the input value. Another solution is to encode the input to adapt the NVM's non-linearity. To the best of our knowledge, this has yet to be studied in detail. The weight update can be written as a sum of outer product between two vectors in many machine learning problems (e.g. stochastic gradient descent, contrastive divergence training of a restricted Boltzmann machine). During the update, write pulses are applied simultaneously across multiple rows and columns. The NVM cells are updated in parallel, with resistance change as a function of the voltages at its corresponding row and column. Different training algorithms exhibit different immunity to weight update non-idealities, and therefore should be studied on a per-case basis. The weight update non-idealities can affect both final training results and training convergence speed. Usually, the deterministic effects such as weight update non-linearity and dynamic range have more impact on the final training accuracy. Stochastic effects such as device-to-device and cycle-to-cycle variations (when not too large) sometimes exhibit correlation with convergence speed. Section 4 provides a literature review of the current state-of-the-art NVM devices used for neuromorphic hardwares in applications ranging from biology based learning models to conventional machine learning algorithms solved using NNs. Sections 5 and 6 provide a more focused overview of the device-level trade-offs required for hardware acceleration of NN architectures using analog in-memory MAC operation.

4. Review of the state-of-the-art devices

This section provides an overview of the emerging NVM technologies that has been utilized as analog synaptic weights in NNs. Inferencing and online learning requires separate set of characteristics from the NVM devices and they will be discussed separately. The desired properties for a NVM device to be used as analog synaptic weight in NNs facilitating MAC operation for inferencing are—large dynamic range of resistance with high (100 kΩ–1 MΩ) value of low resistance state (LRS), high dynamic range of resistance change when programmed with identical pulses in both SET and RESET process, large numbers of distinguishable resistance levels and CMOS logic compatible switching voltage. For online learning, where the weights are updated often, retention is not a big concern but high endurance is desired along with nanoscale switching. For offline inference, where weight is updated occasionally using off-chip learning, good retention characteristics is also required. Any single device has yet to demonstrate all the desired properties. In this paper, we provide a brief review of the three most promising NVM technologies as they are being utilized in neuromorphic applications.

4.1. Resistive random access memory (RRAM)

Among different emerging NVM technolgies, the main advantages of using RRAM for neuromorphic applications, specially for MAC operation for NNs are scalability, moderate switching speed, and low energy consumption. The main challenge for RRAM is to achieve CMOS compatible switching voltage and high endurance. Moreover, the switching, specially the SET operation, is abrupt and makes it difficult to achieve gradual resistivity control by repeated application of the same programming pulse. While it is possible to get gradual RESET operation, RRAMs suffer from non-linearity in switching both during SET and RESET. Also, asymmetry is observed while switching between SET to RESET and RESET to SET. This inherent non-linearity and asymmetry

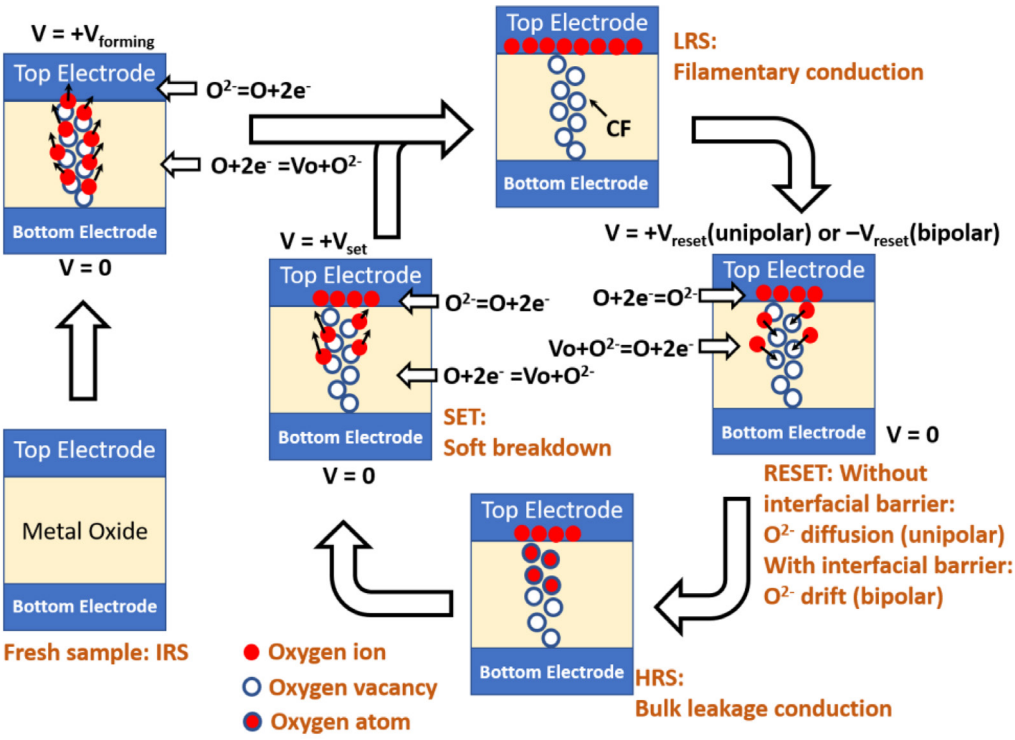

Figure 3. Schematic illustration of the switching process in the simple binary metal-oxide RRAM. [153] © 2012. With permission of Springer.

in the switching of these devices have a negative impact on the accuracy of the NN [49–51]. The other challenging issue in designing NN with RRAM is device-to-device and cycle-to-cycle variation. While some cycle-to-cycle variation can be tolerated in inferencing, it is good to have low device-to-device variation for large arrays. The trade-off between different design constraints and how these impacts the learning and inference accuracy will be discussed in detail in section 5. This section provides an overview of the state-of-the-art RRAM devices utilized for NN application.

Metal oxide RRAM is a simple metal-insulator-metal (MIM) structure, where the insulator layer is typically a transition metal oxide. The metal oxide layer can be a single switching layer or be composed of multiple layers where the interfacial layers are engineered to have the desired properties. Metal oxide RRAMs are also known as valence change memory (VCM), because the resistive switching happens because of the movement of oxygen vacancy defects. These are anion-based memory devices. The other type of RRAM is a cation-based memory device where switching happens because of metal cations diffusion from the anode metal contacts to the solid electrolytes, also known as electrochemical metallization (ECM) cells, will be discussed in the next subsection. These types of cells where the metal cations form a conductive bridge type filament are also termed as 'conductive bridging random access memory (CBRAM)'.

The physics of RRAM devices have been explained by a variety of switching mechanisms, and they have been investigated extensively by the research community [52–58]. The details of the switching mechanism are still an active area of research. The most common switching mechanism is filamentary switching. Here, the set process from the high resistance state (HRS) of the pristine oxide involves soft breakdown of the dielectric material creating a filamentary current conduction path of oxygen vacancy resulting in LRS. The reset process is the switching of the LRS state to the HRS state by recombination of oxygen vacancies with oxygen ions migrated from the electrode/oxide interfacial reservoir upon reversing the bias conditions of the eletrodes as compared to the set state. Figure 3 shows the schematic of the resistive switching mechanism for a binary oxide-RRAM.

For RRAMs to be used in NNs as weight storage, it is often desirable to be able to store analog values, essentially an extreme case of multi-bit operation of a memory, akin to a multi-bit cell (MLC) with many more levels than currently implemented (typically 2- and 3-bit per cell is used for digital non-volatile memories). Numerous RRAM oxide materials have been shown to be capable of multi-bit operation, e.g. Cu_xO [59], TiO_x [60], HfO_x [61], WO_x [62] and TaO_x [63]. One of the early works that demonstrated multi-bit operation was for a TiO_x RRAM [60] where five levels of resistance states was achieved by varying the amplitude of 5 ns voltage pulses. The data retention was 256 h at 85 °C but the endurance was only 2×10^6 cycles. Lee *et al* has also shown five resistance levels without verification for $TiN/TiO_x/HfO_x/TiN$ structure [61]. For the set process multi-level LRS is obtained by changing the set current compliance which modulates the filament diameter or the number of filaments. This compliance dependent multi-level resistance states that results from the modulation of filament size is explained in detail by Chae *et al* [64] and Zhao *et al* [65]. For the reset process, multi-level HRS is obtained by controlling the reset stop voltage. Using Ti

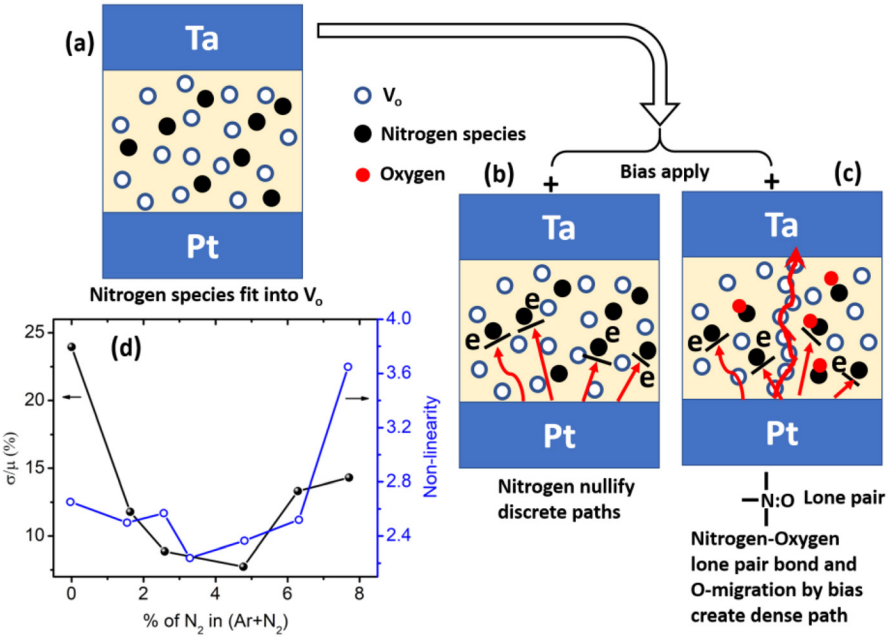

Figure 4. (a)–(c) The schematic representation to describe the denser controlled filament formation by nitrogen incorporation. (d) Analysis of the effect of nitrogen doping on the device of different nitrogen amounts with the function of non-linearity and variability in 30 μA compliance current to set up the guideline for 3-bit MLC storage feasibility of N-TaO_x based RRAM device. Reproduced with permission from [68].

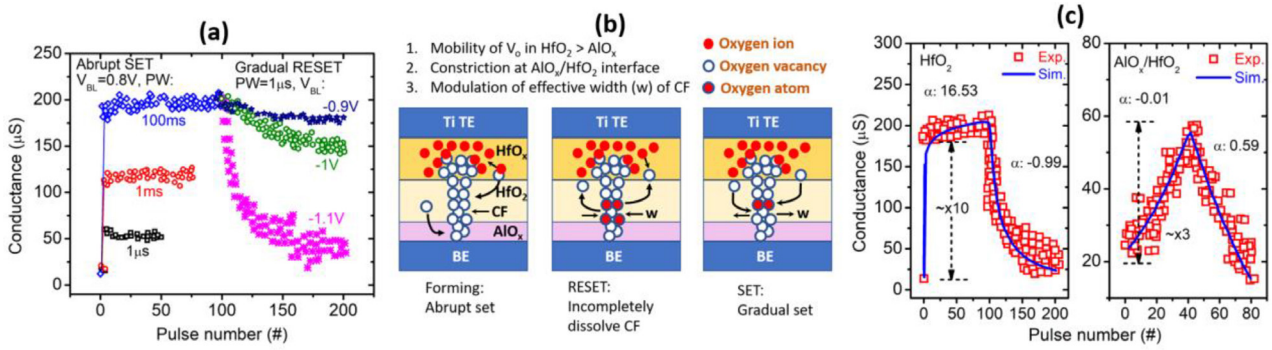

Figure 5. (a) SET and RESET characteristics of the HfO_2 1T-1R array with identical pulses. The conductance change as a function of the number of set/reset pulses is shown. Increment and decrement of the conductance was determined by either a higher voltage or a longer PW. (b) Schematic illustration of an analog switching behavior in the AlO_x/HfO_2 RRAM. (c) Comparison of the SET/RESET switching obtained from the HfO_2 and AlO_x/HfO_2 RRAM devices. In the AlO_x/HfO_2 device, potentiation and depression behavior are obtained by applying identical pulses with 0.9 V and 1 V of 100 μs PW, respectively. © 2016 IEEE. Reprinted, with permission, from [69].

as the oxygen scavenging layer, this structure provides moderately fast operation at 5 ns. The retention is 10 years at 200 °C. While the endurance of 10^6 cycles is enough for training a small dataset as MNIST [66], it is not sufficient for large scale networks with many training examples. Lee *et al* reported one of the highest endurance (10^{12} cycles at 10 ns switching speed) in a memory device made from asymmetric Ta_2O_{5-x}/TaO_{2-x} bilayer structure [67]. Controlling the resistance of the base layer TaO_{2-x} is a means to control the device resistance. However, the switching voltage is rather high ($V_{set} = -4.5$ V, $V_{reset} = +6$ V at 10 ns pulse) and multi-level switching is not reported in this work. Nitrogen doping of TaO_x switching material has been shown to improve multi-bit operation by reducing both the switching voltage and resistance variability [68]. Misha *et al* studied the effect of N doping in TaO_x and reported a device with eight levels of resistive switching [68]. Figure 4 shows the mechanism of nitrogen incorporation in the oxygen vacancy which confines the filament. Nitrogen doping of TaO_x film is reported to reduce the switching variability of voltage and resistance by negating the excess conduction path. This results in the capture of oxygen ion by nitrogen during the bias application that forms the filament confined in a localized region (figure 4(c)). This reduced variability in the filament formation (figure 4(d)) for different compliance current results in higher levels of resistance switching, where the optimized doping results in eight levels of switching with uniform switching among 50 cycles per level.

The SET operation in filamentary RRAM is inherently abrupt in nature. This results in non-linear conductivity switching with the number of switching pulses, which has a negative

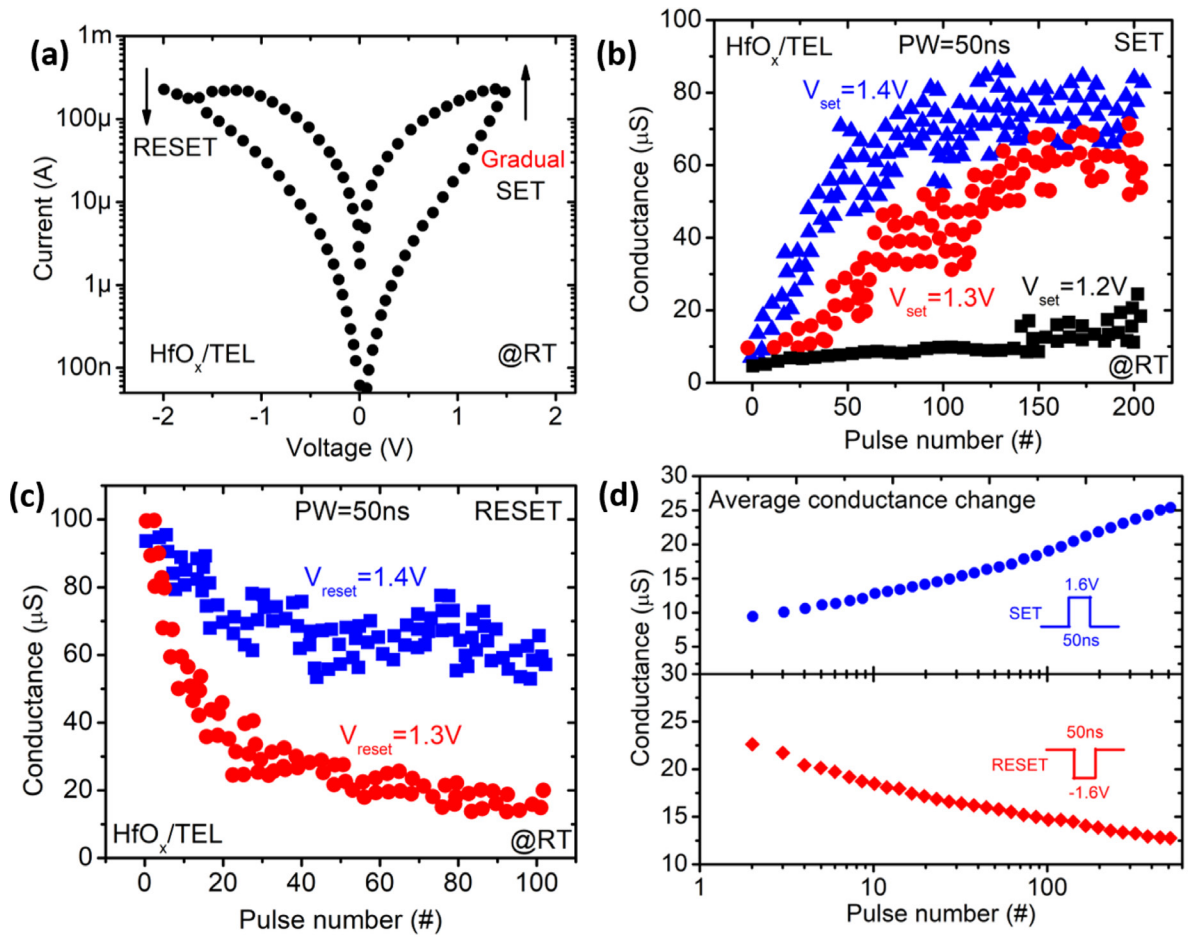

Figure 6. (a) Typical DC-IV of HfO_x/TEL RRAM at room temperature. Analog switching is improved due to TEL layer. (b) Conductance of HfO_x/TEL RRAM changes with number of identical SET pulses at RT. (c) Conductance of HfO_x/TEL RRAM changes with number of identical RESET pulses at RT. (d) Average conductance change during SET and RESET of 256 HfO_x/TEL RRAM devices in the array. © 2017 IEEE. Reprinted, with permission, from [70].

impact on the accuracy of machine learning task. RESET on the other hand is more gradual as shown in figure 5(a) for the TiN (BE)/HfO_2/Ti/TiN (TE) device stack [69]. A barrier layer on the bottom electrode of this device is inserted to avoid an abrupt switching, which resulted in a linear gradual SET/RESET process. Figure 5(c) shows the comparative synaptic behavior observed from a TiN (BE)/HfO_2/Ti/TiN (TE) and an Al (BE)/AlO_x/HfO_2/Ti/TiN (TE) device. In the bilayer system, there is a difference in oxygen vacancy mobility between two layers. During the RESET process, the dissolution of the vacancy is limited by the AlO_x layer because of the lower mobility of oxygen vacancy. Instead the conductance of the conductive filament (CF) is modulated by the width of the filament (figure 5(b)). This results in gradual resistive switching at the expense of low on/off ratio because the width modulation of the filament changes the resistivity according to ohms law compared to the case of tunnel barrier modulation in the length direction, which has an exponential relation with the current. Pattern recognition accuracy increases from 20% for HfO_2 device to close to 90% for the bilayer device.

Wu *et al* proposes that abrupt switching in HfO_x can be explained by the positive feedback of electric field on the formation of CF which accelerates the formation of one single dominant CF [70]. The formation and rupture of one dominant filament contributes to the total conductance change by a significant amount resulting in an abrupt switching behavior. A transition from the abrupt switching to the analog switching is found at higher temperature by confining heat in the switching layer using a thermal enhanced layer (TEL) [70]. Confining heat in the switching layer allows the oxygen vacancies to redistribute themselves uniformly. This results in the formation of multiple weak CFs instead of one dominant filament. This results in a better analog switching behavior as shown in figure 6, where more than ten times of switching window is demonstrated for 50 ns switching pulse.

Amorphous Si (a-Si) barrier layer has been shown to work as an oxygen scavenging layer introducing significant oxygen vacancy in the switching layer (TiO_2) [71]. This results in analog non-filamentary switching with better device to device uniformity than AlO_x barrier layer. However, the switching voltage is relatively large (~6 V) because of relatively thicker a-Si which causes a large voltage drop across it.

Besides material innovation for improved analog switching, 3D device architecture is another important research direction because it provides the advantage of area scaling and increased functionality. 3D vertical RRAM (VRRAM) has

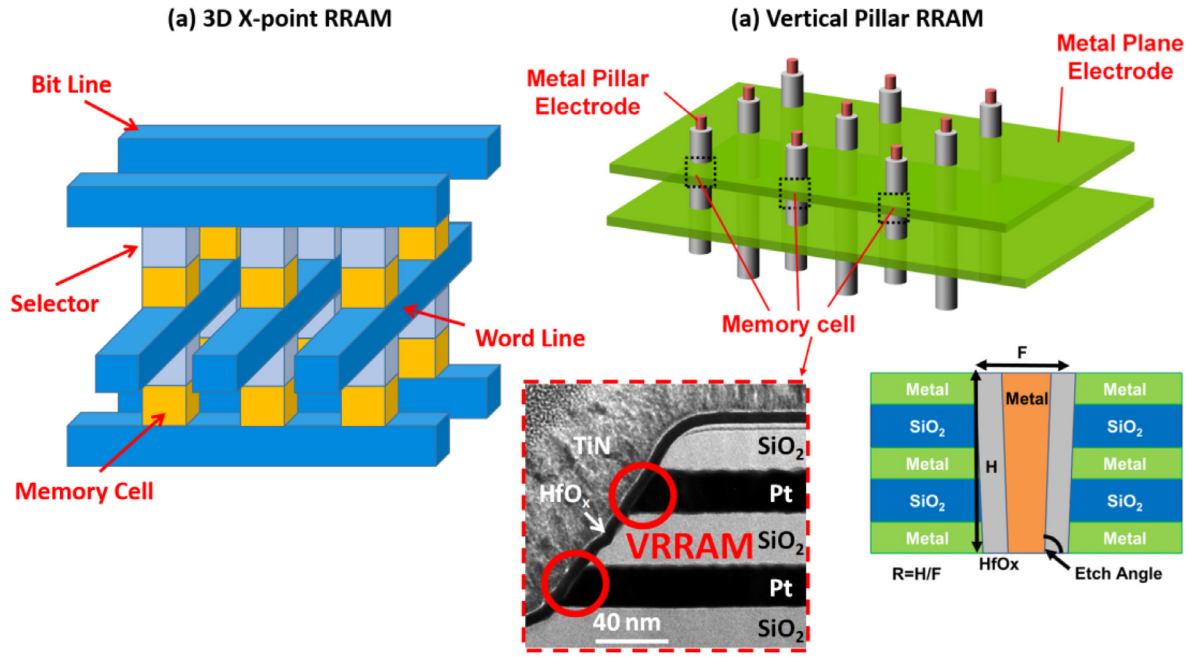

Figure 7. Schematic drawings of (a) 3D X-point™ ReRAM [154], (b) vertical ReRAM. (a) Adapted from [78]. (b) © 2012 IEEE. Reprinted, with permission, from [77].

been demonstrated by several groups (typical structure shown in figure 7) [72–77]. Using 3D VRRAM, Li *et al* introduced a brain-inspired computational framework capable of one-shot learning known as hyperdimensional (HD) computing [78]. Due to the energy efficient VRRAM cells and dense connectivity, this architecture reduces total energy consumption by 52.2% having 412 times less area compared to a low-power digital design using registers as memory. Moreover, this architecture is resilient to RRAM endurance failure because of device-architecture co-optimization.

RRAM arrays have been used successfully for various machine learning tasks. Park *et al* proposed a PCMO (the device stack Pt/AlO_x/TiN_x/$Pr_{0.7}Ca_{0.3}MnO_3$/Pt from top to bottom) based RRAM synaptic device which exhibits the necessary gradual and symmetric conductance change [79]. Using a single layer perceptron of 192 synapses, this device array can learn and recognize human thought pattern corresponding to three vowels from EEG signals. Prezioso *et al* demonstrated transistor-free (1R type) metal oxide RRAM device array crossbars to allow integrated operation of NNs [80]. The bilayer device stack Pt/Ti/TiO_{2-x}/Al_2O_3/Pt is used for an integrated crossbar array of 12 × 12 devices. This single layer perceptron network can be taught to perform the perfect classification of 3 × 3-pixel black and white images into three classes. Gao *et al* demonstrated a convolution kernel operation (i.e. edge detection) on a MNIST image using a 12 × 12 crossbar array with HfO_x RRAM [81]. A recent work by Yao *et al* demonstrated grey-scale human face classification using 128 × 8 array with parallel on-line training [82]. The network designed with optimized metal oxide device stack of TiN/TaO_x/$HfAl_yO_x$/TiN consumes 1000 times less energy than an implementation of the same network using an Intel Xeon Phi processor with an off-chip weight storage. While these demonstrations use NVM as the synaptic device, all of these use circuitry external to the NVM (either in software or in hardware). None of these have NVM integrated with the peripheral control circuits.

4.2. Conductive bridging random access memory (CBRAM)

Filamentary resistive switching devices where the filament is composed of metal cations instead of oxygen vacancies are termed as 'conductive bridging RAM' or CBRAM. The structure of CBRAM devices consists of one electrochemically active electrode (e.g. Ag or Cu that is oxidized easily under an external positive bias) and one electrochemically inert electrode (e.g. Pt, Ir, Au, W, TiN). The switching material between these two electrodes can be a solid electrolyte (chalcogenides) or an oxide material. The first CBRAM-like switching device was proposed by Hirose *et al* [83] in 1976 where switching occurred using a Ag dendrite in a Ag doped As_2S_3 film in a Ag/As_2S_3/Mo structure. Germanium (Ge) based chalcogenide materials ($GeSe_x$ [84], GeS_2 [85], GeTe [86]) have been widely studied as CBRAM active switching material where Cu and Ag ions show high mobility in the chalcogenide materials. The basic mechanism of switching in CBRAM involves electrochemical reaction at the active anode metal (Ag or Cu) which allows metal to form cations. These cations drift through the solid electrolyte switching layer under the electric field and reduces to metal atoms near the inert electrodes. This process forms a metallic conductive bridge from anode to cathode when the device switches from HRS to LRS (SET), hence the name CBRAM. By changing the polarity of the voltage, an electrochemical dissolution of the conductive bridge occurs that resets the device from LRS to HRS. The growth kinetics depend on the electrode and switching materials; therefore, it varies from oxide to non-oxide switching materials. Besides chalcogenides, oxides are widely used for

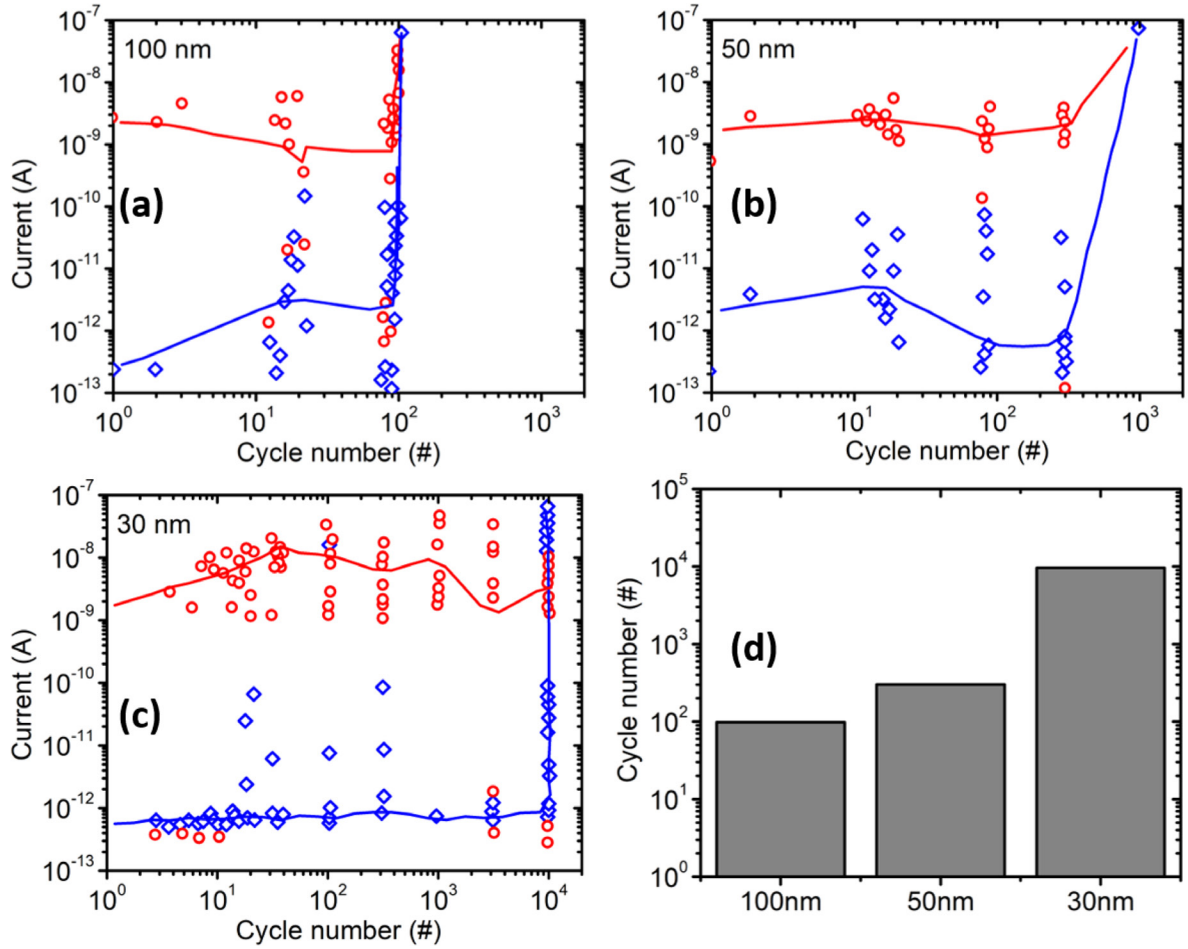

Figure 8. Cycling endurance of the devices with switching layer diameter of 100 nm (a), 50 nm (b), and 30 nm (c). Endurance is improved to 10^4 with scaling down the switching layer area to 30 nm (d). © 2018 IEEE. Reprinted, with permission, from [101].

CBRAM, e.g. SiO_2 [87], ZrO_2 [88], Ta_2O_5 [89], GeO_x [90], TiO_2 [91]. Amorphous Si (a-Si) with Ag doping has also been reported [92]. CBRAM usually has low switching voltage (<2 V), fast switching (~ns), high scalability and low power operation [93, 94]. However, the switching is highly stochastic and abrupt in nature. This creates a challenge for MAC operation in NN where gradual and linear conductivity switching is desirable. Also, achieving high endurance and retention is a challenge. The main reason for these challenges is the highly mobile nature of metal cations for which the diffusion barrier is relatively low in the traditional electrolytes. To control Cu or Ag diffusion to improve switching uniformity, bilayer materials, which create additional cation diffusion barrier, have been studied, e.g. MoO_x/GdO_x [95], Ti/TaO_x [96], $GeSe_x/TaO_x$ [97], $Cu–Te/Al_2O_3$ [98], TiW/Al_2O_3 [99] and so on.

For example, Aratani *et al* demonstrated $>10^7$ cycle endurance from $Cu–Te/GdO_x$ bilayer CBRAM [100]. Four levels of conductive switching were obtained by setting the appropriate compliance current. Precise control of cation injection into the switching layer is the key to improve reliability [100]. Besides the use of a bilayer structure, introducing a transistor in series can also be an effective solution for controlling cation injection. This technique, however, is not suitable for large 2D cross point architecture that is essential for Kirchoff's law type vector matrix multiplier for NN application. Recently, Fujii *et al* demonstrated that confinement of the area of the switching layer in a CBRAM type device is a promising way to control Cu injection [101]. Figure 8 shows that when the SiO_2 switching layer in $Cu/SiO_2/Pt$ CBRAM is reduced from 100 nm to 30 nm in lateral dimension, endurance is improved by two orders of magnitude. The improvement in endurance originates from providing only a limited supply of Cu ions during the set operation due to the spatial limitation of the Cu top electrode. This prevents excessive Cu ions from moving into the SiO_2. It is also reported that reducing the Cu electrode down to sub-20 nm could improve data retention due to the restricted Cu movement within the switching layer.

Using a physical model of the CBRAM, Yu *et al* showed that CBRAMs can emulate the function of a biological synapse, exhibiting STDP behavior, a key observation from biology [102]. One interesting alternative to devices with deterministic multilevel resistance switching is to use devices that show binary switching along with a stochastic-STDP learning rule. This alternative is a functional equivalent with deterministic multilevel synapses at the system-level [103]. Such stochastic binary synapses have been applied to both supervised [104] and unsupervised [105] NN. In this scheme, stochastic switching in resistive memories makes the SET/

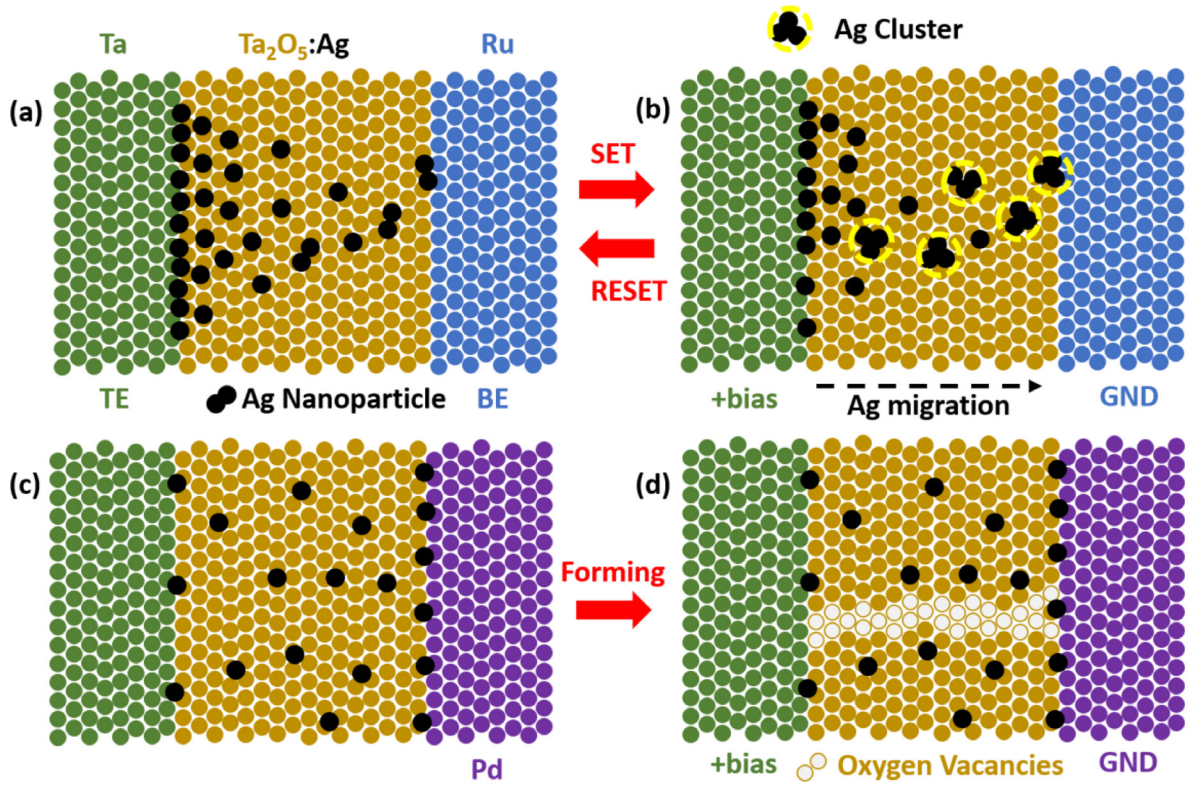

Figure 9. Schematic diagrams illustrate (a) the distributions of Ag ions in the pristine state (equivalent to the HRS) and (b) the LRS (caused by the migration of Ag) in Ta/Ta_2O_5:Ag/Ru device. Schematic diagrams show (c) the pristine state and (d) the forming process (oxygen vacancy mediated VCM) in a Ta/Ta_2O_5:Ag/Pd device. [107] John Wiley & Sons. © 2017 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim.

RESET process probabilistic. The input and the weights of the NN can be converted to a Bernoulli distribution [106] that represents the stochastically switched CBRAM. Formation of one dominant cation filament where the metals have higher diffusivity is the reason for abrupt switching and variability in CBRAM.

Besides stochastic switching, analog resistance modulation based synaptic device using CBRAM has been shown. Jo *et al* proposed a CBRAM device with co-sputtered Ag and Si layer with properly designed Ag/Si mixture ratio gradient that leads to the formation of a Ag-rich (high conductivity) and Ag-poor (low conductivity) region [92]. Ag nanoparticles are embedded into the Si medium that forms a uniform conduction front between Ag-rich and Ag-poor regions. With applied bias, this device shows reliable analog switching behavior having gradual conductance change with subsequent pulses. The analog switching occurs because of the gradual movement of incorporated Ag nanoparticles that allows current conduction through tunneling across Ag nanoparticles as opposed to the formation of a continuous metallic filament. Continuous conductivity modulation as shown in this work for STDP like synaptic operation is also essential for analog weight storage for MAC operation in NN application.

To take the advantage of relatively higher reliability from vacancy-based RRAM along with low voltage operation from CBRAM, Yoon *et al* proposed Ag doped Ta_2O_5 resistive switching device with tantalum (Ta) as the top electrode and ruthenium (Ru) as the bottom electrode [107]. This device does not operate as the traditional CBRAM since the TE does not supply the cation Ag, which remains embedded in the oxide. CMOS compatible switching voltage (0.7 V) is reported with 5×10^7 endurance cycle at 100 ns pulse. The device also shows 9.936×10^6 s retention at room temperature and electro-forming free operation making it one of the most promising devices for neuromorphic application. Ru as BE plays a special role in lowering the switching current and forming free operation compared to a Pd BE as shown in figure 9. There is no mutual solubility between Ag and Ru, resulting in Ru BE repelling Ag atoms away from the BE. This allows Ag to form nanoclusters inside Ta_2O_5 dispersed relatively close to each other resulting in conductive tunneling path (CTP) between the Ag nanoclusters. Unlike CBRAM, there is no continuous cation filament formed here which keeps it forming free. However, in case of Pd BE devices, Ag and Pd can form single uniform phase which makes Ag to be attracted to the BE and get uniformly distributed. This prevents cluster formation. Without the CTP, the switching in Pd BE device is through oxygen vacancies and therefore forming is needed. This work thus exemplifies the need for interface engineering between the electrodes and the switching material to obtain the desired switching performance and the reliability.

4.3. Phase change memory (PCM)

Phase change memories (PCM) are a class of NVM devices where large differences in electrical resistivity between amorphous (high-resistivity) and crystalline (low-resistivity) phases of certain materials are utilized to represent memory

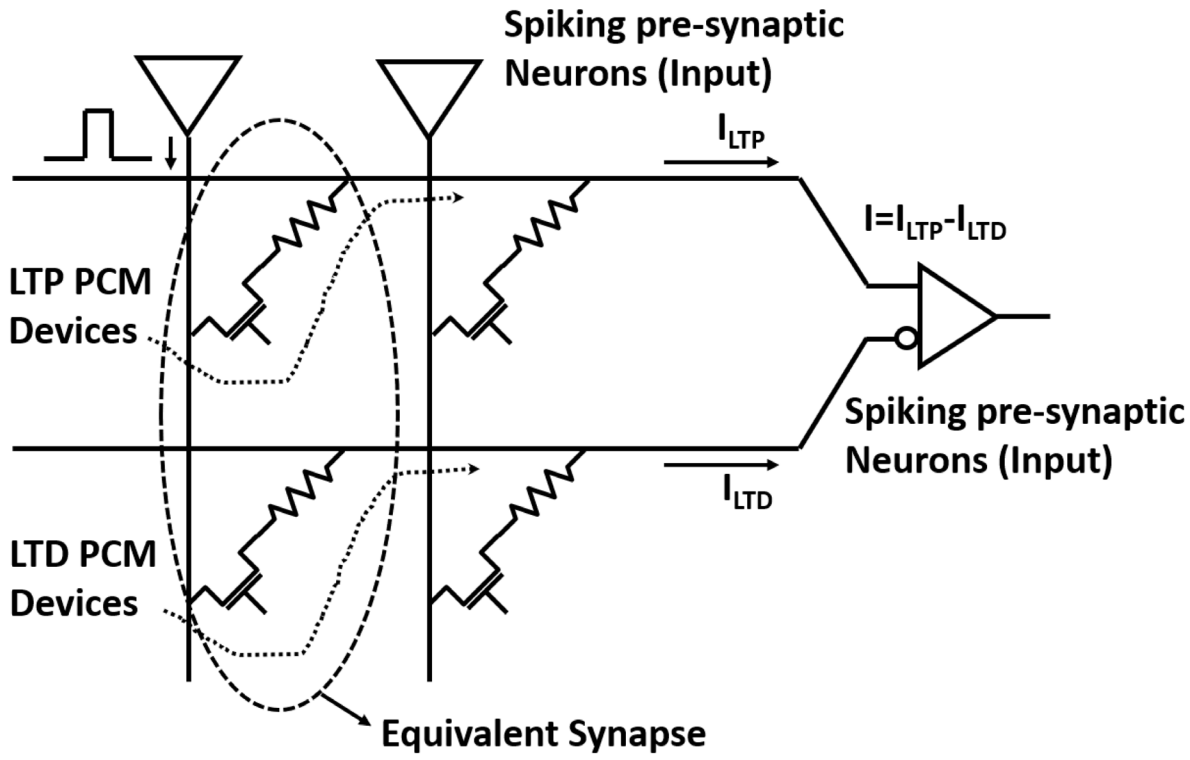

Figure 10. Circuit schematic of the '2-PCM synapse'. Reproduced with permission from [111].

states. The phase transformation occurs through Joule-heating from the current that drives through the phase change material when a voltage pulse is applied. Resistance modulation of phase change materials can also occur by applying voltage pulses with specific amplitude and duration leading to multiple sizes of the amorphous region of the device having resistances between fully amorphous and crystalline state. This behavior enables multiple resistance level operation of PCM, a feature essential for neuromorphic application.

Chalcogenide type materials are widely used in the current PCM technology as phase change materials because of its strong resistance contrast, fast crystallization and high crystallization temperature. More specifically, GST ($Ge_2Sb_2Te_5$), which is located in the pseudo-binary line between GeTe and Sb_2Te_3 in phase diagram, is one of the commonly used materials for memory and synaptic device applications [108]. For PCM devices to be used as a synaptic device, high dynamic range (ratio between high and LRSs) is desired. Since neuromorphic applications also require gradual changes in device resistance with constant voltage pulse, SET process is suitable for this, where repetitive pulse slowly crystallizes high-resistivity amorphous state resulting in a gradual change in resistivity. However, the RESET process is quite abrupt since 'melt and quench' is required for crystalline to amorphous phase transition. Therefore, the SET and RESET resistivity switching for PCM is not symmetric.

PCM is one of the most mature NVM technologies and therefore has gained a lot of interest from the research community as an electronic synapse in neuromorphic computing systems. Kuzum *et al* first demonstrated a single-element phase change electronic synapse with the capability of both the modulation of the time constant and the realization of the different STDP types [109]. Using optical programming, Wright *et al* demonstrated arithmetic operation such as addition, multiplication, subtraction and division in PCM devices [110]. Since amorphization of the phase change material is more abrupt and power consuming than crystallization, Suri *et al* proposed a '2-PCM' synapse circuit where each synapse is represented by 2 PCM devices connected in complementary way to the post-synaptic neuron (figure 10) [111]. One device implements long-term potentiation (LTP, or increase in conductance) and the other device implements long-term depression (LTD, or decrease in conductance), which makes the STDP learning possible using identical crystallization pulses alone. Moreover, the 2-PCM approach also allows us to have both the positive and negative weights. Suri *et al* also improved the synaptic characteristics (SET/RESET current reduction and increase in the number of resistance states) of the standard GST based PCM devices using a thin interfacial layer of HfO_2 which increases the dynamic switching range by improving the crystallization kinetics of the GST film where the interfacial layer can lower the activation energy associated with crystallization and amorphization [112].

The 2-PCM synapse approach has been used by Burr *et al* in backpropagation training for a three-layer perceptron NN. In this network, 164,885 2-PCM synapses were used for vector-matrix multiplication [113]. In an experimental demonstration, Eryilmaz *et al* employed a Hopfield network consisting of 100 synaptic devices and 10 recurrently connected neurons for implementation of brain-like associative learning [114]. Kim *et al* developed a 64k cell (256 × 256) PCM array with on-chip neuron circuits capable of continuous *in situ* learning where a novel 2T1R (two transistors one resistor)

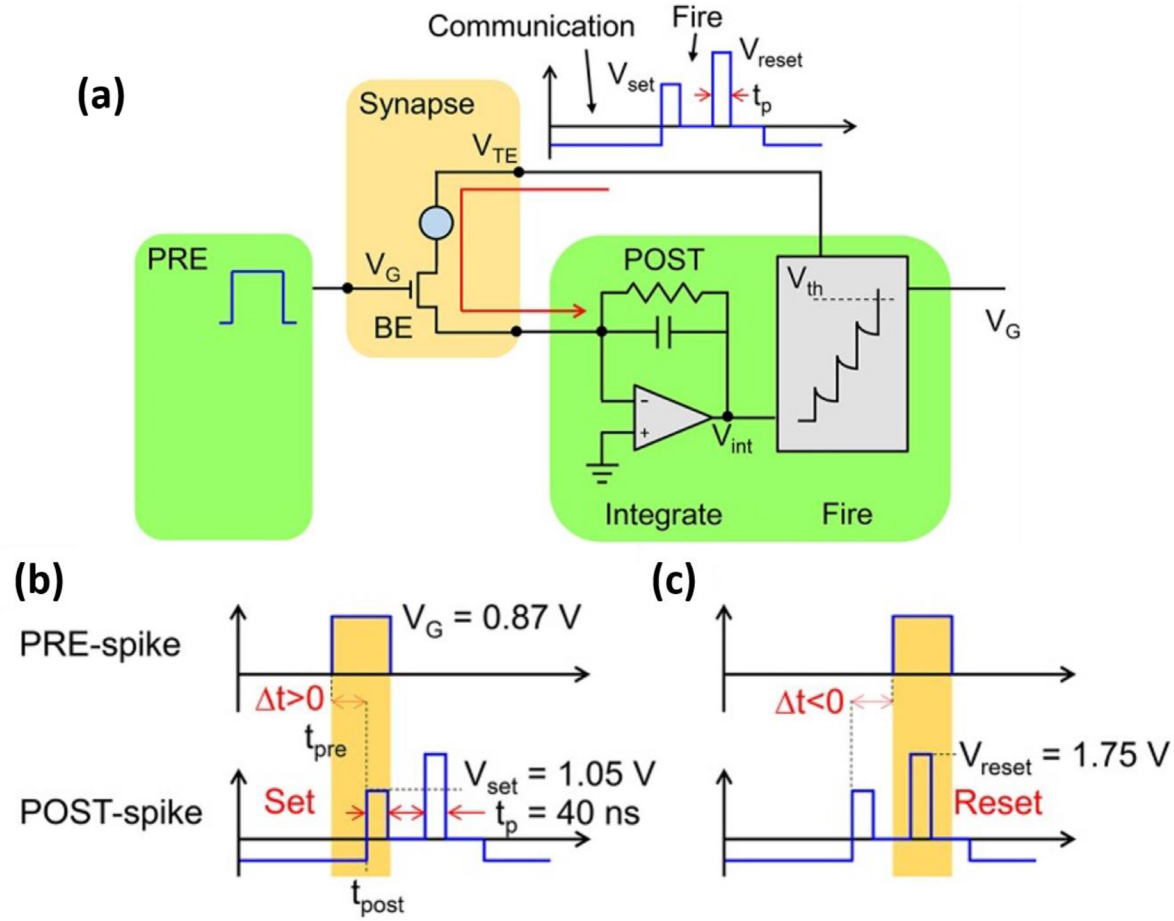

Figure 11. (a) Schematic illustration of the neuromorphic network with a 1T1R synapse. The PRE drives the MOS transistor gate voltage *V*, thus activating a current spike due to the low negative TE voltage ($V_{TE} = 30$ mV) set by the POST. The current spikes are fed into the POST, which eventually delivers a *V* spike back to the synapse as the internal voltage *V* exceeds a threshold *V*. The *V* spike includes a set and reset pulse to induce potentiation/depression according to the STDP protocol. (b) Small positive delay and (c) small negative delay schemes of the applied pulses from the PRE and POST neurons to the 1T1R synapse. Reproduced from [116].

circuit performs both leaky integrate-and-fire (LIF) and STDP learning model asynchronously [115].

Not only supervised learning, but also unsupervised learning has been demonstrated using PCM array. Ambrogio *et al* demonstrated 1T1R PCM synaptic array for unsupervised learning [116]. Using the circuit and pulse scheme shown in figure 11, visual pattern recognition with two or three fully connected neuromorphic layers has been shown with high accuracy (95.5%). Recently, Sebastian *et al* reported that an unsupervised machine-learning algorithm, running on one million PCM devices, successfully found temporal correlations in unknown data streams [117]. This work uses the linear resistance switching property of the multi-level memory device to solve linear differential equation. These devices utilize PCM crystallization dynamics to perform both computation (detecting temporal correlations between event-based data streams) and storage of the results and can be considered as 'computational memory devices'. Application of different NVM devices for various neuromorphic applications require trade-offs in device performance and reliability metric. The next section will discuss the topic in detail.

The highlights of section 4 are summarized in table 1 and figure 12.

5. Design trade-off in NVM devices for different NN applications

5.1. Retention and endurance

To capture the correlation between electrical parameters of the synaptic device and microscopic factors and to investigate the intrinsic trade-off between different parameters, researchers have developed different Monte Carlo simulation methods for both filamentary and non-filamentary RRAM devices [118, 119]. The simulation by Gao *et al* [118, 119] calculates the distribution of electric field, current density and temperature in the local region of the device, where the resistive switching occurs. Then the probability of generation/migration/recovery of the ions or vacancies can be calculated. The calculation is followed by a stochastic dynamic update of the distribution of ions or vacancies. Based on the calculated distribution and evolution of ions or vacancies, the device parameters can also be calculated that can predict the device characteristics.

Table 1. Comparison between different reported RRAM and CBRAM devices with regards to the key device parameters.

	Material stack	Switching voltage (V) (SET/RESET)	Switching levels (Ω) (LRS/HRS)	On/off ratio	Speed (ns)	Retention	Endurance	Reference
RRAM	TiN/TiO_x/HfO_x/TiN	+0.8/−0.8	1 k/1 M	10^3	5	10 years	10^5	[61]
	Pt/Ta_2O_{5-x}/TaO_{2-x}/Pt	−1/+2	30 k/NR	NR	10	10 years @ 85 °C	10^{12}	[67]
	Ta/Ta_2O_5:Ag/Ru	+0.7/−0.7	100 k/10 M	10^2	100	115 d @ RT	5×10^7	[107]
	TiN/ N:HfO_2/Pt	+1/−1	1 k/10 k	10	900	10^4 s @ 85 °C	10^9	[140]
	TiN/TiO_2/a-Si/TiN	+7/−7	<1 μA current for 30 nm device	NR	10	3 years @ 55 °C	10^6	[141]
	Pt/Ti/TiO_{2-x}/Al_2O_3/Pt	−2/+2	10 k/100 k	10	5×10^5	10 years	5×10^3	[80]
	TiN/HfO_x/AlO_x/Pt	+1.4 → +1.8/−2.2 → −2.6	10 k/1 M	10^2	50	7200 s	10^5	[69]
CBRAM	Pt/GeSO/TiN	+0.7/−1.1	200/500	2.5	100	NR	2×10^3	[137]
	TE/Cu–Te/GdO_x/BE	+3/−1.7	10 k/10 M	10^3	5	10^3 s	10^7	[100]
	TE/Ag + Si/Si/BE	3.2/−2.8	25 M/200 M	8	3×10^5	NR	10^7	[92]
	Cu/SiO_2/Pt	+1/−0.5	500 M/5 G	10	NR	NR	10^4	[101]
	Cu/Ta_2O_5/Pt	+3.5/−2.5	100/100 M	10^6	10^4	NR	10^4	[89]
	Cu/TiW/Al_2O_3/W	+1/−1	100 k/100 M	10^3	10	600 s @ 125 °C	10^6	[99]

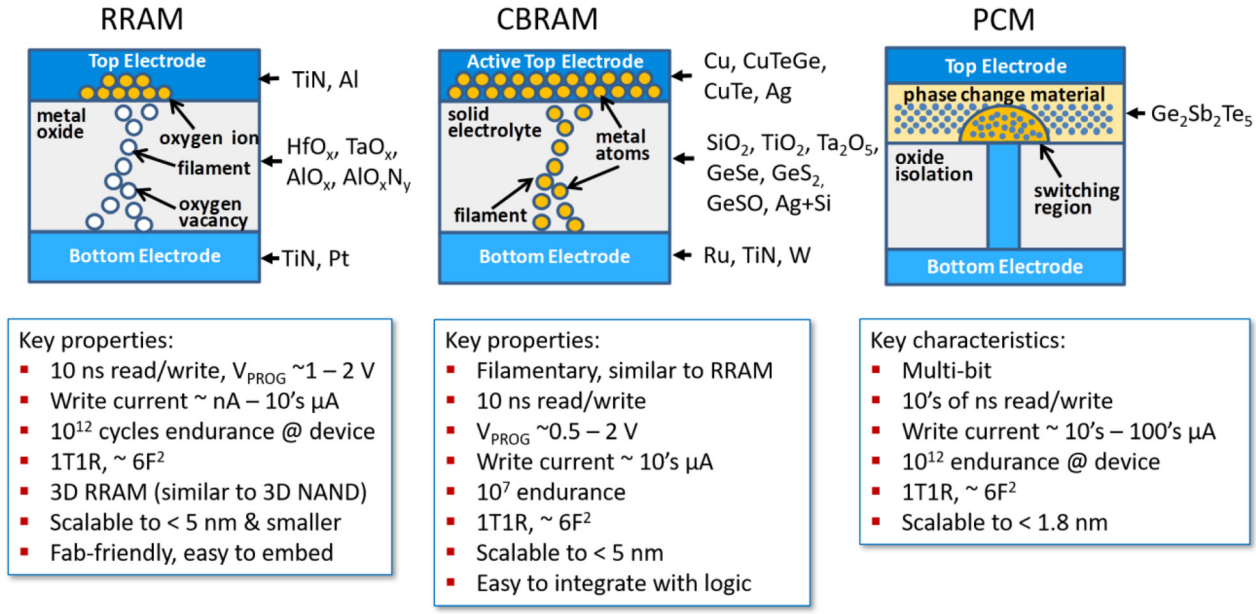

Figure 12. Comparison of RRAM, CBRAM and PCM technology.

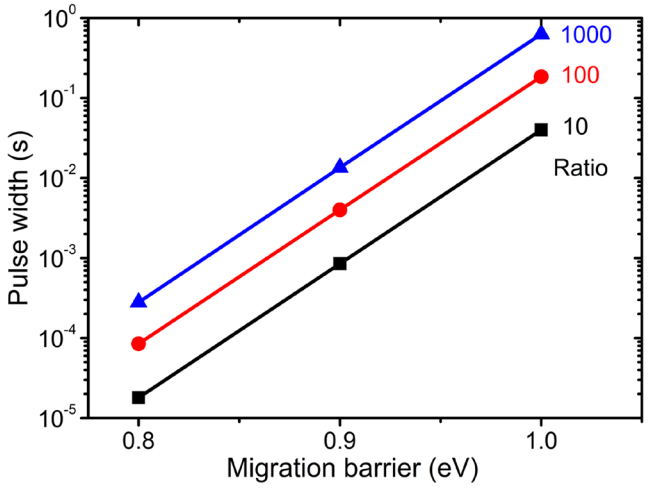

Figure 13. Simulated retention behavior of non-filamentary RRAM devices with various migration barrier of oxygen vacancy. Simulation parameters can be found in [118].

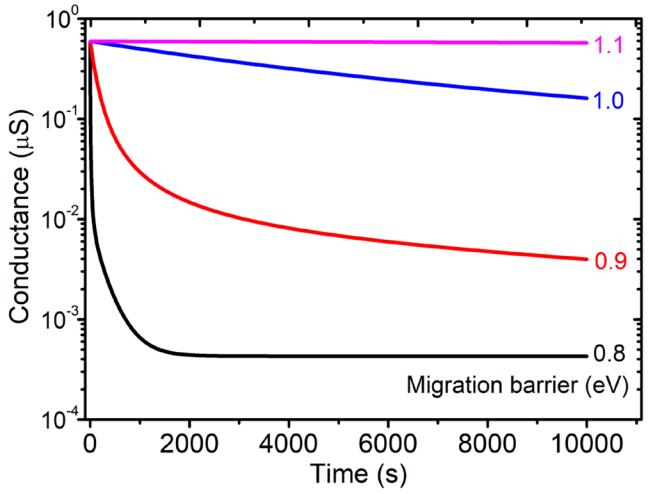

Figure 14. Simulated potentiation process of non-filamentary RRAM devices with various migration barrier of oxygen vacancy. The programming voltage is fixed as 2 V and pulse number is fixed as 100. To program to a larger ratio, longer pulse width is required.

Non-filamentary RRAM devices, also known as interface switching RRAM devices, are suitable for bi-directional analog switching, but they usually suffer from the retention and speed trade-off [118]. The resistive switching of non-filamentary RRAM is attributed to the change of an interfacial electronic barrier modulated by oxygen vacancy migration. As shown in figures 13 and 14, if the migration barrier of oxygen vacancy is higher, the device is more stable, but also requires more time for programming. In contrast, with a lower migration barrier of oxygen vacancy, the programming speed can be increased, but retention degrades very fast. For most of the cases, HRS is the stable state and the resistance of the lower resistance state increases with time. Since non-filamentary RRAM devices were mostly used as analog synapses for online training [73, 80, 120], the research community has aimed at increasing programming speed without considering retention. Even so, the programming speed was still on the order of micro-second, and the reported references on data retention for multilevel states at high temperature were limited. For PCRAM, the trade-off between programming speed and retention can be achieved by modulating the stoichiometry of the GST material with tungsten dopant [121] or applying a constant voltage via prestructural ordering (incubation) effects [122].

On the other hand, filamentary RRAM devices (including both OxRAM and CBRAM), which have widely been investigated for the use as a digital nonvolatile memory, can have both nano-second programming speed and excellent high temperature retention. This is because the programming process of filamentary RRAM originates from oxygen vacancy generation or oxygen interstitial migration, while the retention

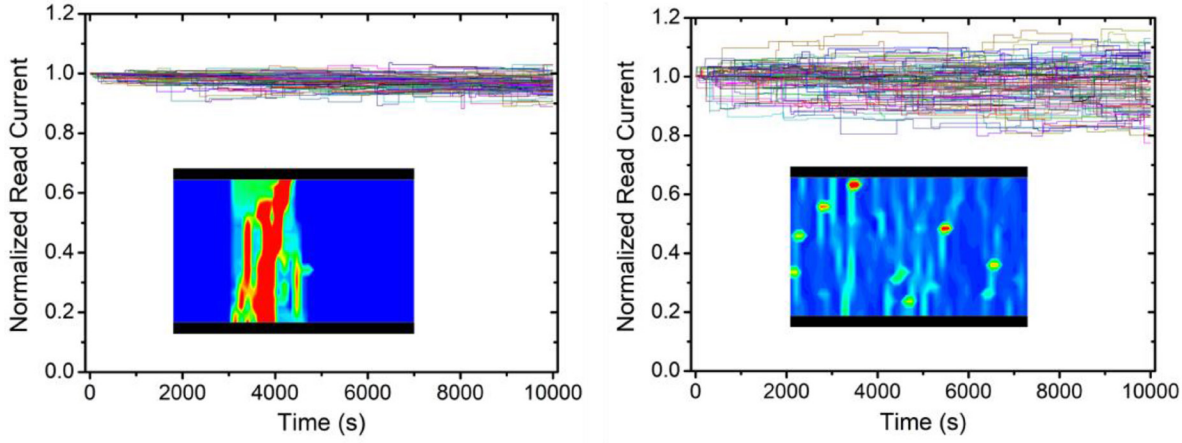

Figure 15. (a) Simulated retention behavior of filamentary RRAM devices with single strong CF. 100 devices from the same original state are simulated and shown. Baking temperature is 85 °C. Simulation details and other parameters can be found in [9]. Inset: current density distribution of the RRAM device. Its order parameter is 0.67. (b) Simulated retention behavior of filamentary RRAM devices with multiple weak CF. The other situation is the same to (a). Inset is the current density, and its order parameter is 0.46. Retention becomes worse in this case.

degradation is due to the oxygen vacancy diffusion [57, 123]. These different mechanisms have different activation energies and obviate the intrinsic trade-off of retention and speed. However, filamentary RRAM faces another trade-off problem between retention and multilevel switching. In CBRAM, the source of the filament is metal ion interstitial migration. So, the aforementioned conclusions are similar. The only difference may be that the activation energy of metal ions is smaller than that of oxygen vacancies; so the CBRAMs are faster but the retention is worse.

Generally, the connection and rupture of the CF causes abrupt resistance change, so filamentary RRAM devices are best utilized as binary synapse [124] or single-bit NVM. Muraoka and Ninomiya *et al* proposed a method to make the oxygen vacancies distributing more tightly, forming a single strong CF with high oxygen vacancy concentration [123, 125]. With this optimization method, oxygen vacancies are not easy to diffuse out from the CF region, and even though some of these oxygen vacancies diffuse out, only a small resistance change will be observed. In this case, retention can be improved significantly. Recently, Gao *et al* proposed that analog switching behaviors could be realized on filamentary RRAM separating the oxygen vacancies to different location, forming multiple weak CFs [119]. Each CF only contributes to a small portion of the total conductance of the device. These CFs are not so stable as the CFs in single-CF device, and thus retention degradation can be observed at high temperature. A similar idea of weak CF was demonstrated in CBRAM using Ag doped SiO_2 [92].

An order parameter was introduced to quantify the distribution of oxygen vacancy [119]. The order parameter is defined as the percentage of vacancy–vacancy neighbored pairs in the whole lattice of switching oxide layer. It can be expressed as $O_V = 2N_{V-V}/zC_VN$, where N_{V-V} is the number of vacancy–vacancy neighbored pairs, C_V is the concentration of oxygen vacancy, N is the total number of oxygen sites in the oxide layer, and z is the coordinate number of lattice. As shown in figure 15, if the order parameter is large (ordered state), which means a strong CF is formed and the device cannot show good analog switching, the retention is high. Whereas, if the order parameter is small (disordered state), which means the device is designed for good analog switching and multiple weak CFs may be formed, resistance fluctuation is observed under high temperature baking. To improve the retention, doping method or multi-layered structure were developed to avoid oxygen vacancy diffusion from its original location [126]. However, doping will introduce discrete dopant variations when the device is scaled down to a smaller size.

Endurance is another key parameter for device reliability. Till now, there are few works reporting the endurance of analog switching NVM. For binary switching, which was mainly aimed for use as digital memory, degradation of endurance were extensively investigated [126]. Chen *et al* found that there is a tradeoff between endurance and retention [127]. To get better endurance, the oxygen reservoir layer is very important. This layer could control the concentration of oxygen vacancy in the resistive switching layer, avoiding quick loss of oxygen ions. Besides retention and endurance, read disturbance is another impportant reliability parameter [128]. In a NN, read disturbance dictates how many times of inference process the network can do without refreshing the weights. Continuous reading may change resistance state of the devices and degrade the accuracy of the network. Although there has been no clear conclusion, it is widely accepted that read disturbance is correlated with the retention degradation, and somewhat analogous to a voltage accelerated retention degradation process [129, 130].

5.2. Operating voltage

Reducing the operating voltage is important for lowering the power consumption of the NN. Specifically, an operating voltage of less than 1 V is essential for CMOS-compatible on-chip integration of neuromorphic devices. The read voltage may

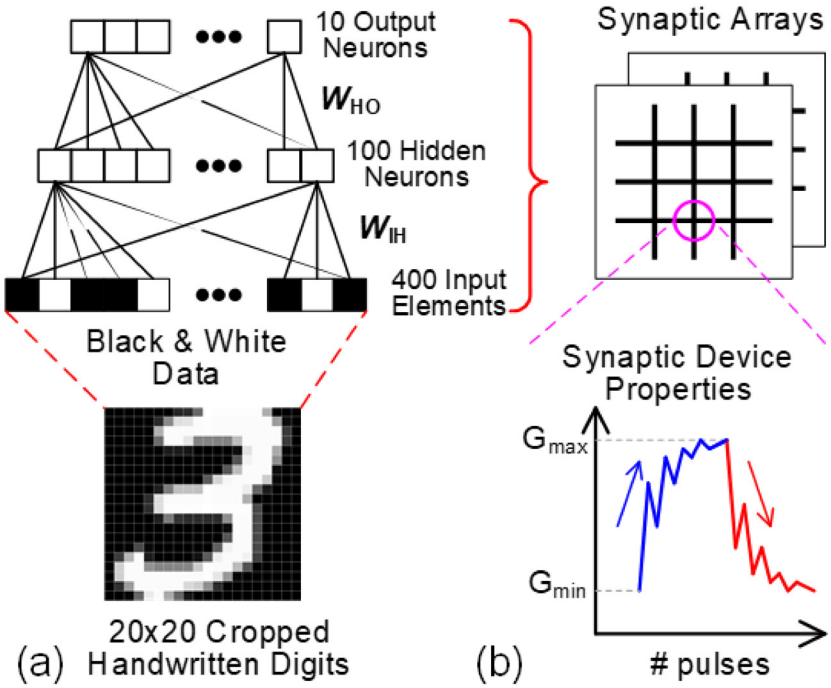

Figure 16. (a) The 2-layer MLP NN. The input MNIST images are cropped and encoded into black/white data for simplification. (b) In the MLP simulator, the weights WIH and WHO are implemented with synaptic arrays, where each synaptic device exhibits non-ideal device properties.

also determine the total scale of the network since larger read voltages result in larger read current. In a highly parallel NN, the large read current through bit lines may limit the array size. If the non-linearity of *I–V* curve is very large, increasing read voltage will significantly increase the read current. However, due to current fluctuation, the read voltage cannot decrease too much. The current fluctuation mainly comes from the random telegraph noise caused by electron trapping/detrapping and oxygen (ions or vacancies) vibration [119, 131, 132]. Typically, only one or several traps or oxygen vacancies contribute to the current fluctuation, so the amplitude of current fluctuation is almost independent of the current level. With a small read voltage, current fluctuation contributes a large portion to the read current and may affect the accuracy of the NN. Therefore, to make the read current more stable, read voltage should be kept at a reasonable range and cannot be too small.

The programming voltage depends on the SET/RESET voltage of the synaptic devices. It should be higher than the threshold (SET/RESET voltage) for switching and cannot be too large to avoid hard breakdown. The voltage-time-dilemma indicates that reducing programming voltage linearly will incur an exponential increase in programming time [53, 57]. The SET/RESET voltage only depends on the synaptic device itself and is usually less than 2 V. But sometimes a barrier layer is designed for nonlinear *I–V* curve or better reliability. The new layer may take up part of the applied voltage and thus increase the SET/RESET voltage by up to several volts. Meanwhile, it should also be noticed that too small SET/RESET voltage may cause a read disturbance issue [130]. If the read voltage is close to the SET/RESET voltage of the device, the resistance may change very fast during the inference process. This discussion is valid for both RRAM and CBRAM.

5.3. Resistance levels and variability

To study the feasibility of synaptic devices as analog weights on NNs, a simulator (NeuroSim) has been developed [133] for a 2-layer MLP NN with synaptic device properties incorporated into the weights. As shown in figure 16, MNIST handwritten digits are used [134] as the training and testing dataset to implement online learning and offline classification. The MLP network topology is 400 (input layer)–100 (hidden layer)–10 (output layer). 400 neurons of the input layer correspond to 20 × 20 MNIST image (converted to black/white and edge cropped), and 10 neurons of output layer correspond to 10 classes of digits. Such a simple 2-layer MLP can achieve 96%–97% in the software baseline. In online learning, the MLP simulator takes into account the synaptic device properties in training the network with images randomly picked from the training dataset (60 k images) and classifying the testing dataset (10 k images). In offline classification, the network is pre-trained by software, and the MLP simulator only performs the classification with synaptic device properties.

As shown in figure 17, several non-ideal synaptic device properties in the simulator is evaluated such as non-linear and noisy weight update, limited weight precision and finite weight range, etc. To analyze the effect of nonlinear weight update, a set of nonlinear curves are defined and labeled with nonlinearity values from 6 to −6 for both the potentiation (weight increase) and depression (weight decrease). The potentiation and depression will not necessarily follow the same trajectory due to the non-linearity of weight update, resulting in the asymmetry with positive non-linearity value for potentiation and negative nonlinearity for depression. Experiments performed by various groups show that the potentiation

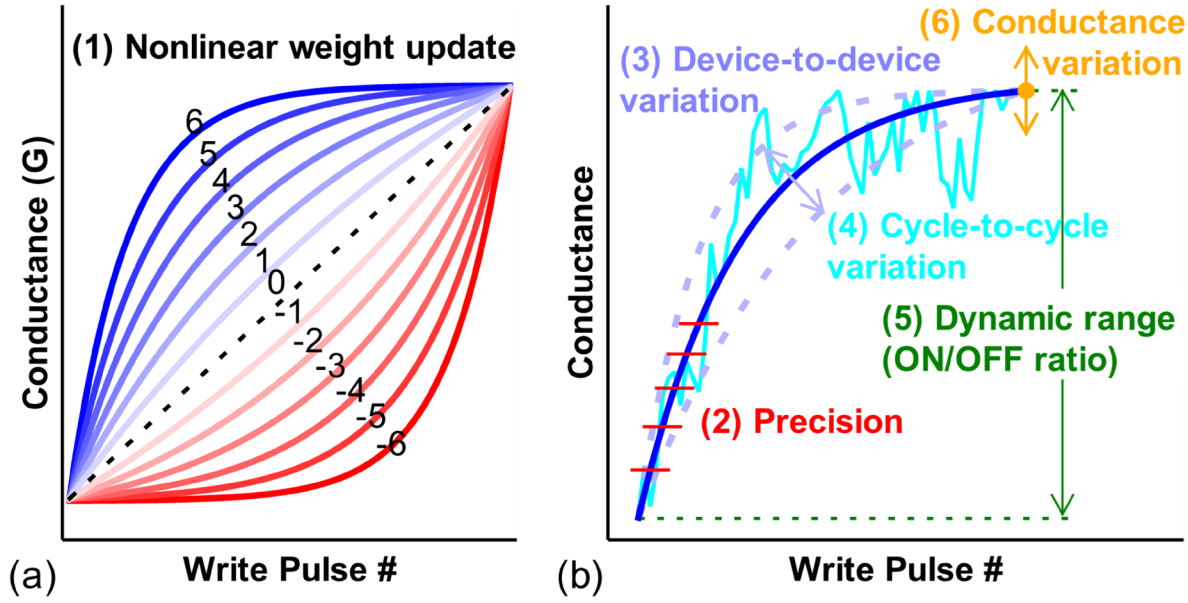

Figure 17. Schematic illustration of non-ideal synaptic device properties modeled in the MLP simulator, including (1) nonlinear weight update (a), (2) weight precision, (3) device-to-device weight update variation, (4) cycle-to-cycle weight update variation, (5) dynamic range (conductance ON/OFF ratio) and (6) conductance variation (b).

and the depression have positive and negative nonlinearity, respectively [69, 92, 120, 135]. During the weight update, the device's conductance is tuned within a confined conductance range, and only a finite number of conductance states are available due to the weight precision. Ideally, the lowest conductance state (OFF state) should be low enough to represent the zero weight in the algorithm, making the dynamic range (conductance ON/OFF ratio) sufficiently large. In reality, the ON/OFF ratio is always finite and normally not large enough. Different devices may even observe different ON/OFF ratios if the conductance range has a variation. On top of the nonlinear weight update curves, there are also considerable weight update variations from device to device, and even from pulse to pulse within one device. The effect of device-to-device weight update variation can be analyzed by introducing the variation into the nonlinearity baseline for each synaptic device, while the cycle-to-cycle variation refers to as the variation in conductance change at every programming pulse.

To quantify the impact of the aforementioned non-ideal device properties, sensitivity analyses was performed for online learning and offline classification using the simulator. Figure 18(a) shows the requirement of weight precision. The result suggests that 6-bit weight is required for online learning, while 2-bit weight is needed for offline classification (at least for MNIST dataset) and 1-bit weight introduces only a slight degradation. Figure 18(b) shows the learning accuracy with different ON/OFF ratios. Limited ON/OFF ratio (<50) will degrade the accuracy of offline classification. The network may adapt itself to this limited ON/OFF ratio during learning thus the online learning can tolerate more ON/OFF ratio (>10 is needed). However, the accuracy drop in online learning is sharper, which is probably because the network will deviate more from its correct form with both erroneous weighted sum and weight update results. Figure 18(c) shows the impact of weight update non-linearity and asymmetry. The result shows that the asymmetry (positive potentiation P and negative depression D) is the key factor that degrades the accuracy, and high non-linearity can be tolerated if P/D have the same polarity. However, for common situations where P/D is positive/negative, the impact of nonlinearity on the online learning accuracy is very critical. High accuracy can only be achieved with small nonlinearity (<1). For offline classification, there is no asymmetry/nonlinearity issue as the cell conductance can be iteratively programmed to the desired value [136]. Variation sensitivity analyses are performed with different asymmetry and non-linearity values (P/D: positive/negative) in online learning. Figure 18(d) shows the impact of conductance range variation on the learning accuracy. We added the variation (with standard deviation (σ) in terms of percentage) on the highest conductance state (ON state) as it changes the conductance range most. The result shows that the conductance variation does not degrade the learning accuracy. Instead, it remedies the accuracy loss due to high nonlinearity. However, an opposite trend can be observed for the device-to-device variation, as shown in figure 18(e).

The amount of device-to-device variation is defined as the standard deviation (σ) of nonlinearity. At low non-linearity (<1), the accuracy slightly decreases with larger variation. For the non-linearity >1, the impact becomes much more prominent. On the other hand, the amount of cycle-to-cycle variation (σ) is expressed in terms of the percentage of the entire weight range. As shown in figure 18(f), small cycle-to-cycle variation ($<2\%$) can alleviate the degradation of learning accuracy by high non-linearity. The reason may be attributed to the random disturbance that aids the convergence of the weights to an optimal weight pattern (i.e. to help the system jump out of local minima). Thus, synaptic devices with non-linear weight update behavior may perform better than expected if they exhibit a little noisy weight update. However, too large variation ($>2\%$) overwhelms the

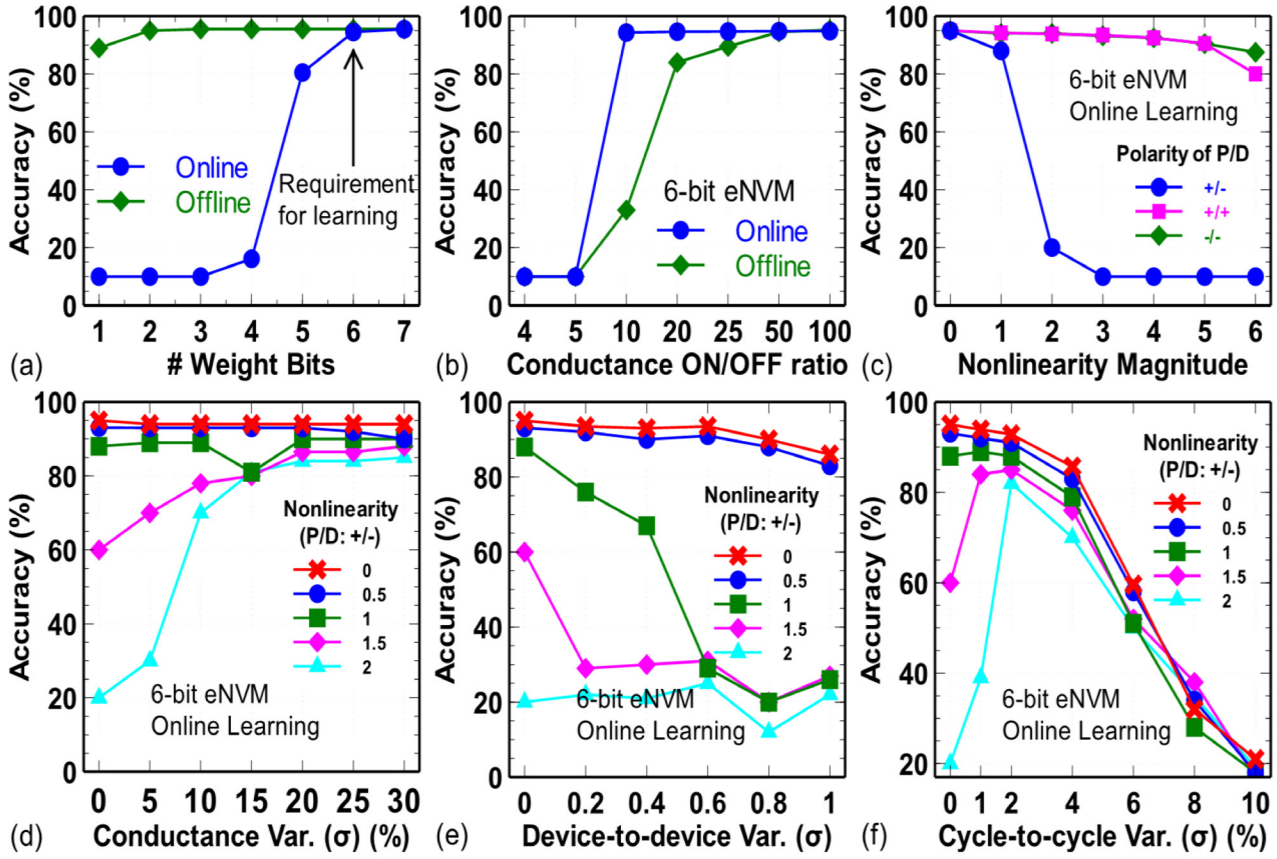

Figure 18. The impact of (a) weight precision, (b) conductance ON/OFF ratio, (c) weight update asymmetry/nonlinearity, (d) conductance range variation, (e) device-to-device variation and (f) cycle-to-cycle variation in online learning and/or offline classification. Reproduced with permission from [133].

deterministic weight update amount defined by the algorithm and thus is harmful to the learning accuracy. This set of simulations help to define the desired synaptic device characteristics that enables high online learning accuracy. To summarize, a symmetric and close to linear weight update with sufficient ON/OFF ratio is critical, while a reasonable amount of device-to-device, cycle-to-cycle variations could be tolerated. As the simulation presented in this section is generalized and based on by varying the device properties, the analysis is technology agnostic and the conclusions are valid for any type of resistive memory devices.

6. Perspective on the device parameters for large scale NN architectures

For a broad class of neuromorphic applications, large conductance switching range with linear response for identical switching pulse are desired. There exists an exponential relationship between switching pulse witdh and pulse amplitude. Low energy switching requirement stipulates that the switching pulse width and pulse amplitude be as small as possible. There has been considerable work done so far in finding the right material combination for the desirable switching characteristics. Based on our review of such devices in sections 4 and 5, we have summarized the current state-of-the-art device parameters reported for neuromorphic application in figure 19. Figure 19(a) shows the switching pulse width as a function of switching pulse amplitude for different RRAM and CBRAM device stacks used for neuromorphic application. Also, figure 19(b) shows the reported conductance range as a function of the pulse amplitude for the corresponding devices. Devices ideally suited to the neuromorphic application should provide large conductance switching range at low pulse amplitude and small pulse width. In figure 19(a), the direction of smaller switching energy is marked with an arrow. The ideal device stack will lie at the corner pointed by the direction of the arrow shown in this figure. Based on this metric, Pt/GeSO/TiN [137], TiN/TaO_x/Pt [138] and TiN/SiO_x/TaO_x/Pt [138] would have been the better choices. But figure 18(b) suggests that these devices show high conductance which is not desirable since a large array of such devices would draw large currents. Also, the range of conductance change is very low. Considering both the figures of merit, the optimum choice would be TiN/HfO_x/AlO_x/Pt [139] (data point 9 in figure 18) which shows two orders of conductance switching at short switching pulse width. Another promising device is HfO_x device with thermally enhanced layer (TEL) [70] (data point 14 in figure 18) which ensures fast switching at low voltage. The conductance also is not too high. The range of conductance switching needs to be increased in order to ensure higher precision matrix-vector multiplication for NN application. Cycle-to-cycle variation limits the number of resistive switching states that also decreases the precision of the matrix-vector multiplication. Simulation suggests that smaller

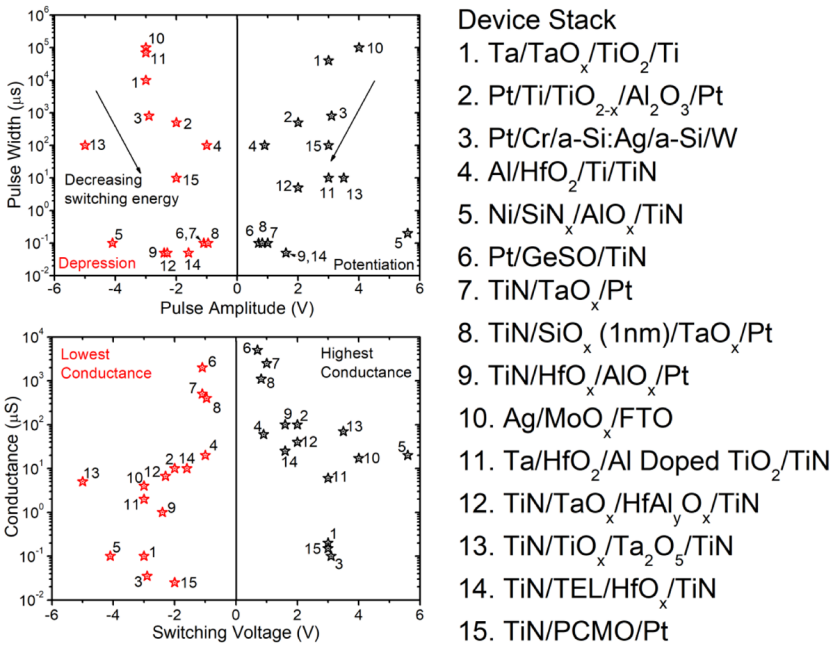

Figure 19. (a) Pulse width versus pulse amplitude for gradual conductance switching for RRAMs and CBRAMs demonstrated in literature. (b) Conductance range of the aforementioned devices during the switching as a function of the pulse amplitude. The device data are taken from [69, 70, 80, 82, 92, 135, 137–139, 155–159].

networks can tolerate some device-to-device variation, but in order to scale up the network lower device-to-device variation is desired. One useful capability of NVM array for in-memory computing is 'blind weight update' which saves additional read during the write sequences by not requiring write-verify scheme. To have such capability in an array, besides low cycle-to-cycle variation, highly linear resistance switching response as a function of identical pulses is required. While this is a limitation for inorganic devices, certain organic devices show high linearity [19]. Inorganic devices with TEL show a lot of promise in this regard [70].

7. Conclusion

The inference and training of today's state-of-the-art DNNs demand extreme energy efficiency beyond general-purpose architecture. General purpose computing architecture cannot provide the optimized dataflow needed to achieve the desired computing throughput at low energy cost for DNNs. Design of specialized hardware accelerator improves the energy efficiency of DNN inference and training by optimizing memory hierarchy and data-flow design, improving parallelism, and leveraging special properties of NNs such as error-tolerance and sparsity. The use of emerging on-chip NVM provides a path for further improving energy efficiency by performing highly-parallel analog multiply-accumulate and weight update directly inside memory and eliminating data movement. The capability to integrate tera-byte scale memory on chip enables hardware design to keep up with the increasing model size and computation complexity of DNN models. On-chip integration of memory provides with highly parallel and high bandwidth, memory access. The inference and training of DNN pose different sets of requirements on NVM device charactersitices. In general, larger conductance range, more intermediate states, and higher resistance are desirable for both inference and traning. For inference, an ideal device should also have linear *I–V* relationship and long retention time. For training, symmetric and linear pulse response, small device-to-device and cycle-to-cyle variation, and good endurance are crucial. In this paper, we reviewed the state-of-the-art emerging NVM devices. None of the devices we have reviewed could simultaneously combine all these favorable properties. Besides further device engineering, it is crucial for hardware designers to select proper material stacks and make reasonable tradeoffs depending on the target application.

The switching mechanism in RRAM involves oxygen ion movement to and from oxygen vacancies. Therefore, controlling the oxygen ion movement during pulsed switching in RRAM can be a promising way to achieve the aforementioned performance goals. Placing an oxygen ion barrier to make a bilayer RRAM and confinement of the generated heat during switching have shown significant improvement in analog switching. Better thermal management in RRAM can also provide filament stability that could improve reliability like retention and endurance. If the ideal device characteristics can be achieved, the most important aspect of Kirchoff's law based analog matrix-vector multiplication array using NVMs is that it can provide ultra-low energy, high throughput computing without compromising bit precision that is currently missing in the neural network accelerator landscape.

Acknowledgments

This work is supported in part by ASCENT (one of six centers in JUMP, a Semiconductor Research Corporation (SRC) program sponsored by DARPA), and member companies

of the Stanford SystemX Alliance and the Stanford NVM Technology Research Initiative (NMTRI). The authors also acknowledge Beijing Innovation Center for Future Chips (ICFC), Beijing Municipal Science and Technology Project (Z181100003218001), and NSFC (61674089, 61674092).

ORCID iDs

Raisul Islam https://orcid.org/0000-0002-1222-6117

References

[1] Mead C 1990 Neuromorphic electronic systems *Proc. IEEE* **78** 1629–36
[2] Horowitz M 2014 1.1 computing's energy problem (and what we can do about it) *IEEE Int. Solid-State Circuits Conf. Digest of Technical Papers* pp 10–14
[3] Ando K *et al* 2017 BRein memory: a 13-layer 4.2 K neuron/0.8 M synapse binary/ternary reconfigurable in-memory deep neural network accelerator in 65 nm CMOS *Symp. on VLSI Circuits* pp C24–5
[4] Tang T, Xia L, Li B, Wang Y and Yang H 2017 Binary convolutional neural network on RRAM *22nd Asia and South Pacific Design Automation Conf.* pp 782–7
[5] Chen Y *et al* 2017 DaDianNao: a machine-learning supercomputer *47th Annual IEEE/ACM Int. Symp. on Microarchitecture* pp 609–22
[6] Han S *et al* 2016 EIE: efficient inference engine on compressed deep neural network *Proc. 43rd Int. Symp. on Computer Architecture* pp 243–54
[7] Chen Y, Krishna T, Emer J S and Sze V 2017 Eyeriss: an energy-efficient reconfigurable accelerator for deep convolutional neural networks *IEEE J. Solid-State Circuits* **52** 127–38
[8] Shafiee A *et al* 2016 ISAAC: a convolutional neural network accelerator with *in situ* analog arithmetic in crossbars *ACM/IEEE 43rd Annual Int. Symp. on Computer Architecture* pp 14–26
[9] Chi P *et al* 2016 PRIME: a novel processing-in-memory architecture for neural network computation in ReRAM-based main memory *ACM/IEEE 43rd Annual Int. Symp. on Computer Architecture* pp 27–39
[10] Angizi S, He Z, Parveen F and Fan D 2017 RIMPA: a new reconfigurable dual-mode in-memory processing architecture with spin Hall effect-driven domain wall motion device *IEEE Computer Society Annual Symp. on VLSI* pp 45–50
[11] Parashar A *et al* 2017 SCNN: an accelerator for compressed-sparse convolutional neural networks *Proc. 44th Annual Int. Symp. on Computer Architecture* pp 27–40
[12] Fan D and Angizi S 2017 Energy efficient in-memory binary deep neural network accelerator with dual-mode SOT-MRAM *IEEE Int. Conf. on Computer Design* pp 609–12
[13] Andri R, Cavigelli L, Rossi D and Benini L 2018 YodaNN: an architecture for ultralow power binary-weight CNN acceleration *IEEE Trans. Comput. Des. Integr. Circuits Syst.* **37** 48–60
[14] Eryilmaz S B, Kuzum D, Yu S and Wong H S P 2015 Device and system level design considerations for analog-non-volatile-memory based neuromorphic architectures *Technical Digest—International Electron Devices Meeting*
[15] Kuzum D, Yu S and Philip Wong H S 2013 Synaptic electronics: materials, devices and applications *Nanotechnology* **24** 382001
[16] Chua L O 1971 Memristor—the missing circuit element *IEEE Trans. Circuit Theory* **18** 507–19
[17] Strukov D B, Snider G S, Stewart D R and Williams R S 2008 The missing memristor found *Nature* **453** 80–3
[18] Vongehr S and Meng X 2015 The missing memristor has not been found *Sci. Rep.* **5** 11657
[19] van de Burgt Y, Melianas A, Keene S T, Malliaras G and Salleo A 2018 Organic electronics for neuromorphic computing *Nat. Electron.* **1** 386–97
[20] Burr G W *et al* 2017 Neuromorphic computing using non-volatile memory *Adv. Phys.* X **2** 89–124
[21] Yu S 2018 Neuro-inspired computing with emerging nonvolatile memorys *Proc. IEEE* **106** 260–85
[22] Hartley M, Taylor N and Taylor J 2006 Understanding spike-time-dependent plasticity: a biologically motivated computational model *Neurocomputing* **69** 2005–16
[23] He K, Zhang X, Ren S and Sun J 2016 Deep residual learning for image recognition *IEEE Conf. on Computer Vision and Pattern Recognition*
[24] Krizhevsky A, Sutskever I and Hinton G E 2012 ImageNet classification with deep convolutional neural networks *Proc. 25th Int. Conf. on Neural Information Processing Systems* pp 1097–105
[25] LeCun Y, Bengio Y and Hinton G 2015 Deep learning *Nature* **521** 436
[26] Simonyan K and Zisserman A 2014 Very deep convolutional networks for large-scale image recognition (arXiv:1409.1556v6)
[27] Szegedy C *et al* 2015 Going deeper with convolutions *IEEE Conf. on Computer Vision and Pattern Recognition* pp 1–9
[28] Karpathy A, Toderici G, Shetty S, Leung T, Sukthankar R and Fei-Fei L 2014 Large-scale video classification with convolutional neural networks *Proc. IEEE Conf. on Computer Vision and Pattern Recognition* pp 1725–32
[29] Intel™ architecture instruction set extensions programming reference (https://software.intel.com/sites/default/files/managed/c5/15/architecture-instruction-set-extensions-programming-reference.pdf)
[30] Durant L, Giroux O, Harris M and Stam N 2017 Inside volta: the world's most advanced data center GPU *NVIDIA Developer Blog*
[31] CNN-benchmarks (https://github.com/jcjohnson/cnn-benchmarks)
[32] Gokhale V, Jin J, Dundar A, Martini B and Culurciello E 2014 A 240 G-ops/s mobile coprocessor for deep neural networks *IEEE Conf. on Computer Vision and Pattern Recognition Workshops* pp 696–701
[33] Chen T *et al* 2014 DianNao: a small-footprint high-throughput accelerator for ubiquitous machine-learning *Proc. 19th Int. Conf. on Architectural Support for Programming Languages and Operating Systems* pp 269–84
[34] Du Z *et al* 2015 ShiDianNao: Shifting vision processing closer to the sensor *ACM/IEEE 42nd Annual Int. Symp. on Computer Architecture* pp 92–104
[35] Zhang C, Li P, Sun G, Guan Y, Xiao B and Cong J 2015 Optimizing FPGA-based accelerator design for deep convolutional neural networks *Proc. 2015 ACM/SIGDA Int. Symp. on Field-Programmable Gate Arrays* pp 161–70
[36] Park S, Bong K, Shin D, Lee J, Choi S and Yoo H 2015 4.6 A1.93TOPS/W scalable deep learning/inference processor with tetra-parallel MIMD architecture for big-data applications *IEEE Int. Solid-State Circuits Conf. Digest of Technical Papers* pp 1–3
[37] Jouppi N P *et al* 2017 In-datacenter performance analysis of a tensor processing unit *Proc. 44th Annual Int. Symp. on Computer Architecture* pp 1–12
[38] Moons B and Verhelst M 2016 A 0.3–2.6 TOPS/W precision-scalable processor for real-time large-scale ConvNets *IEEE Symp. on VLSI Circuits* pp 1–2

[39] Redmon J, Divvala S K, Girshick R B and Farhadi A 2015 You only look once: unified, real-time object detection (arXiv:1506.02640v5)

[40] Natarajan S *et al* 2014 A 14 nm logic technology featuring 2nd-generation FinFET, air-gapped interconnects, self-aligned double patterning and a 0.0588 μm^2 SRAM cell size *IEEE Int. Electron Devices Meeting* pp 3.7.1–3

[41] Wu S *et al* 2016 A 7 nm CMOS platform technology featuring 4th generation FinFET transistors with a 0.027 μm^2 high density 6-T SRAM cell for mobile SoC applications *IEEE Int. Electron Devices Meeting* pp 2.6.1–4

[42] Nvidia volta architecture (http://images.nvidia.com/content/volta-architecture/pdf/volta-architecture-whitepaper.pdf)

[43] Villa C, Mills D, Barkley G, Giduturi H, Schippers S and Vimercati D 2010 A 45 nm 1 Gb 1.8 V phase-change memory *IEEE Int. Solid-State Circuits Conf.* pp 270–1

[44] Choi Y *et al* 2012 A 20 nm 1.8 V 8 Gb PRAM with 40 MB s^{-1} program bandwidth *IEEE Int. Solid-State Circuits Conf.* pp 46–8

[45] Fackenthal R *et al* 2014 19.7 A 16 Gb ReRAM with 200 MB s^{-1} write and 1 GB s^{-1} read in 27 nm technology *IEEE Int. Solid-State Circuits Conf. Digest of Technical Papers* pp 338–9

[46] Liu T *et al* 2014 A 130.7 mm^2 2-Layer 32 Gb ReRAM memory device in 24 nm technology *IEEE J. Solid-State Circuits* **49** 140–53

[47] Otsuka W *et al* 2011 A 4 Mb conductive-bridge resistive memory with 2.3 GB s^{-1} read-throughput and 216 MB s^{-1} program-throughput *IEEE Int. Solid-State Circuits Conf.* pp 210–1

[48] Rho K *et al* 2017 23.5 A 4 Gb LPDDR2 STT-MRAM with compact 9F2 1T1MTJ cell and hierarchical bitline architecture *IEEE Int. Solid-State Circuits Conf.* pp 396–7

[49] Sidler S *et al* 2016 Large-scale neural networks implemented with non-volatile memory as the synaptic weight element: impact of conductance response *46th European Solid-State Device Research Conf.* pp 440–3

[50] Chen P, Gao L and Yu S 2016 Design of resistive synaptic array for implementing on-chip sparse learning *IEEE Trans. Multi-Scale Comput. Syst.* **2** 257–64

[51] Chen P Y, Li Z and Yu S 2016 Design tradeoffs of vertical RRAM-based 3D cross-point array *IEEE Trans. Very Large Scale Integr. Syst.* **24** 3460–7

[52] Waser R 2009 Resistive non-volatile memory devices *Microelectron. Eng.* **86** 1925–8

[53] Waser R, Dittmann R, Staikov G and Szot K 2009 Redox-based resistive switching memories—nanoionic mechanisms, prospects, and challenges *Adv. Mater.* **21** 2632–63

[54] Waser R and Aono M 2007 Nanoionics-based resistive switching memories *Nat. Mater.* **6** 833

[55] Sawa A 2008 Resistive switching in transition metal oxides *Mater. Today* **11** 28–36

[56] Akinaga H and Shima H 2010 Resistive random access memory (ReRAM) based on metal oxides *Proc. IEEE* **98** 2237–51

[57] Wong H-S P *et al* 2012 Metal-oxide RRAM *Proc. IEEE* **100** 1951–70

[58] Jeong D S *et al* 2012 Emerging memories: resistive switching mechanisms and current status *Rep. Prog. Phys.* **75** 076502

[59] Wang S Y, Huang C W, Lee D Y, Tseng T Y and Chang T C 2010 Multilevel resistive switching in Ti/Cu_xO/Pt memory devices *J. Appl. Phys.* **108** 114110

[60] Yoshida C, Tsunoda K, Noshiro H and Sugiyama Y 2007 High speed resistive switching in Pt/TiO_2/TiN film for nonvolatile memory application *Appl. Phys. Lett.* **91** 223510

[61] Lee H Y *et al* 2008 Low power and high speed bipolar switching with a thin reactive Ti buffer layer in robust HfO_2 based RRAM *IEEE Int. Electron Devices Meeting* pp 1–4

[62] Lai E K *et al* 2010 Tungsten oxide resistive memory using rapid thermal oxidation of tungsten plugs *Japan. J. Appl. Phys.* **49** 04DD17

[63] Terai M, Sakotsubo Y, Kotsuji S and Hada H 2010 Resistance controllability of Ta_2O_5/TiO_2 stack ReRAM for low-voltage and multilevel operation *IEEE Electron Device Lett.* **31** 204–6

[64] Chae B G *et al* 2017 Nanometer-scale phase transformation determines threshold and memory switching mechanism *Adv. Mater.* **29** 1701752

[65] Zhao X, Xu H, Wang Z, Zhang L, Ma J and Liu Y 2015 Nonvolatile/volatile behaviors and quantized conductance observed in resistive switching memory based on amorphous carbon *Carbon* **91** 38–44

[66] Yu S *et al* 2017 Binary neural network with 16 Mb RRAM macro chip for classification and online training *Technical Digest—Int. Electron Devices Meeting*

[67] Lee M J *et al* 2011 A fast, high-endurance and scalable non-volatile memory device made from asymmetric Ta_2O_{5-x}/TaO_{2-x} bilayer structures *Nat. Mater.* **10** 625–30

[68] Misha S H *et al* 2015 Effect of nitrogen doping on variability of TaO_x -RRAM for low-power 3-bit MLC applications *ECS Solid State Lett.* **4** P25–8

[69] Woo J *et al* 2016 Improved synaptic behavior under identical pulses using AlO_x/HfO_2 bilayer RRAM array for neuromorphic systems *IEEE Electron Device Lett.* **37** 994–7

[70] Wu W, Wu H, Gao B, Deng N, Yu S and Qian H 2017 Improving analog switching in HfO_x based resistive memory with thermal enhanced layer *IEEE Electron Device Lett.* **38** 1019–22

[71] Govoreanu B *et al* 2016 Advanced a-VMCO resistive switching memory through inner interface engineering with wide ($<10^2$) on/off window, tunable μA-range switching current and excellent variability *IEEE Symp. on VLSI Technology* pp 1–2

[72] Bai Y *et al* 2015 Stacked 3D RRAM array with graphene/CNT as edge electrodes *Sci. Rep.* **5** 13785

[73] Wang I-T, Lin Y-C, Wang Y-F, Hsu C-W and Hou T-H 2014 3D synaptic architecture with ultralow sub-10 fJ energy per spike for neuromorphic computation *IEEE Int. Electron Devices Meeting* pp 28.5.1–4

[74] Sohn J, Lee S, Jiang Z, Chen H and Wong H-P 2014 Atomically thin graphene plane electrode for 3D RRAM *IEEE Int. Electron Devices Meeting* pp 5.3.1–4

[75] Park S *et al* 2012 A non-linear ReRAM cell with sub-1 μA ultralow operating current for high density vertical resistive memory (VRRAM) *Int. Electron Devices Meeting* pp 20.8.1–4

[76] Yu M *et al* 2016 Novel vertical 3D structure of TaO_x-based RRAM with self-localized switching region by sidewall electrode oxidation *Sci. Rep.* **6** 21020

[77] Chen H-Y, Yu S, Gao B, Huang P, Kang J and Wong H S P 2012 HfO_x based vertical resistive random access memory for cost-effective 3D cross-point architecture without cell selector *Int. Electron Devices Meeting* pp 20.7.1–4

[78] Li H *et al* 2016 Hyperdimensional computing with 3D VRRAM in-memory kernels: device-architecture co-design for energy-efficient, error-resilient language recognition *IEEE Int. Electron Devices Meeting* pp 16.1.1–4

[79] Park S *et al* 2015 Electronic system with memristive synapses for pattern recognition *Sci. Rep.* **5** 10123

[80] Prezioso M, Merrikh-Bayat F, Hoskins B D, Adam G C, Likharev K K and Strukov D B 2015 Training and operation of an integrated neuromorphic network based on metal-oxide memristors *Nature* **521** 61–4

[81] Gao L, Chen P and Yu S 2016 Demonstration of convolution kernel operation on resistive cross-point array *IEEE Electron Device Lett.* **37** 870–3

[82] Yao P *et al* 2017 Face classification using electronic synapses *Nat. Commun.* **8** 15199

[83] Hirose Y and Hirose H 1976 Polarity-dependent memory switching and behavior of Ag dendrite in Ag-photodoped amorphous As_2S_3 films *J. Appl. Phys.* **47** 2767–72

[84] Rahaman S Z *et al* 2012 Enhanced nanoscale resistive switching memory characteristics and switching mechanism using high-Ge-content $Ge_{0.5}Se_{0.5}$ solid electrolyte *Nanoscale Res. Lett.* **7** 614

[85] Vianello E *et al* 2012 Sb-doped GeS_2 as performance and reliability booster in conductive bridge RAM *Int. Electron Devices Meeting* pp 31.5.1–4

[86] Choi S, Lee J, Bae H, Yang W, Kim T and Kim K 2009 Improvement of CBRAM resistance window by scaling down electrode size in pure-GeTe film *IEEE Electron Device Lett.* **30** 120–2

[87] Schindler C, Thermadam S C P, Waser R and Kozicki M N 2007 Bipolar and unipolar resistive switching in Cu-doped SiO_2 *IEEE Trans. Electron Devices* **54** 2762–8

[88] Li Y *et al* 2010 Resistive switching properties of Au/ZrO_2/Ag structure for low-voltage nonvolatile memory applications *IEEE Electron Device Lett.* **31** 117–9

[89] Banno N *et al* 2008 Diffusivity of Cu ions in solid electrolyte and its effect on the performance of nanometer-scale switch *IEEE Trans. Electron Devices* **55** 3283–7

[90] Rahaman S Z *et al* 2012 Repeatable unipolar/bipolar resistive memory characteristics and switching mechanism using a Cu nanofilament in a GeO_x film *Appl. Phys. Lett.* **101** 73106

[91] Tada M *et al* 2009 Highly scalable nonvolatile TiO_x/$TaSiO_y$ solid-electrolyte crossbar switch integrated in local interconnect for low power reconfigurable logic *IEEE Int. Electron Devices Meeting* pp 1–4

[92] Jo S H, Chang T, Ebong I, Bhadviya B B, Mazumder P and Lu W 2010 Nanoscale memristor device as synapse in neuromorphic systems *Nano Lett.* **10** 1297–301

[93] Valov I, Waser R, Jameson J R and Kozicki M N 2011 Electrochemical metallization memories—fundamentals, applications, prospects *Nanotechnology* **22** 254003

[94] Jana D *et al* 2015 Conductive-bridging random access memory: challenges and opportunity for 3D architecture *Nanoscale Res. Lett.* **10** 188

[95] Yoon J *et al* 2009 Excellent switching uniformity of Cu-doped MoO_x/GdO_x bilayer for nonvolatile memory applications *IEEE Electron Device Lett.* **30** 457–9

[96] Rahaman S Z *et al* 2012 Excellent resistive memory characteristics and switching mechanism using a Ti nanolayer at the Cu/TaO_x interface *Nanoscale Res. Lett.* **7** 345

[97] Rahaman S Z *et al* 2012 Impact of TaO_x nanolayer at the $GeSe_x$/W interface on resistive switching memory performance and investigation of Cu nanofilament *J. Appl. Phys.* **111** 63710

[98] Goux L *et al* 2011 Influence of the Cu–Te composition and microstructure on the resistive switching of Cu–Te/Al_2O_3/Si cells *Appl. Phys. Lett.* **99** 53502

[99] Belmonte A *et al* 2013 90 nm W\Al_2O_3\TiW\Cu 1T1R CBRAM cell showing low-power, fast and disturb-free operation *5th IEEE Int. Memory Workshop* pp 26–9

[100] Aratani K *et al* 2007 A novel resistance memory with high scalability and nanosecond switching *IEEE Int. Electron Devices Meeting* pp 783–6

[101] Fujii S *et al* 2018 Scaling the CBRAM switching layer diameter to 30 nm improves cycling endurance *IEEE Electron Device Lett.* **39** 23–6

[102] Yu S and Wong H-P 2010 Modeling the switching dynamics of programmable-metallization-cell (PMC) memory and its application as synapse device for a neuromorphic computation system *Int. Electron Devices Meeting* pp 22.1.1–4

[103] Neftci E O, Pedroni B U, Joshi S, Al-Shedivat M and Cauwenberghs G 2016 Stochastic synapses enable efficient brain-inspired learning machines *Frontier Neurosci.* **10** 241

[104] Lee J H and Likharev K K 2007 Defect-tolerant nanoelectronic pattern classifiers *Int. J. Circuit Theory Appl.* **35** 239–64

[105] Suri M *et al* 2013 Bio-inspired stochastic computing using binary CBRAM synapses *IEEE Trans. Electron Devices* **60** 2402–9

[106] Merkel C, Kudithipudi D, Suri M and Wysocki B 2017 Stochastic CBRAM-based neuromorphic time series prediction system *J. Emerg. Technol. Comput. Syst.* **13** 37:1–14

[107] Yoon J H *et al* 2017 Truly electroforming-free and low-energy memristors with preconditioned conductive tunneling paths *Adv. Funct. Mater.* **27** 1702010

[108] Shi Y, Fong S, Wong H S P and Kuzum D 2014 Synaptic devices based on phase-change memory *Neuro-Inspired Computing Using Resistive Synaptic Devices* ed S Yu (Cham: Springer) pp 19–51

[109] Kuzum D, Jeyasingh R G D, Lee B and Wong H S P 2012 Nanoelectronic programmable synapses based on phase change materials for brain-inspired computing *Nano Lett.* **12** 2179–86

[110] Wright C D, Liu Y, Kohary K I, Aziz M M and Hicken R J 2011 Arithmetic and biologically-inspired computing using phase-change materials *Adv. Mater.* **23** 3408–13

[111] Suri M *et al* 2011 Phase change memory as synapse for ultra-dense neuromorphic systems: application to complex visual pattern extraction *Int. Electron Devices Meeting* pp 4.4.1–4

[112] Suri M *et al* 2012 Interface engineering of PCM for improved synaptic performance in neuromorphic systems *4th IEEE Int. Memory Workshop* pp 1–4

[113] Burr G W *et al* 2014 Experimental demonstration and tolerancing of a large-scale neural network (165 000 synapses), using phase-change memory as the synaptic weight element *IEEE Int. Electron Devices Meeting* pp 29.5.1–4

[114] Eryilmaz S B *et al* 2013 Experimental demonstration of array-level learning with phase change synaptic devices *IEEE Int. Electron Devices Meeting* pp 25.5.1–4

[115] Kim S *et al* 2015 NVM neuromorphic core with 64 k-cell (256-by-256) phase change memory synaptic array with on-chip neuron circuits for continuous *in situ* learning *IEEE Int. Electron Devices Meeting* pp 17.1.1–4

[116] Ambrogio S *et al* 2016 Unsupervised learning by spike timing dependent plasticity in phase change memory (PCM) synapses *Frontiers Neurosci.* **10** 56

[117] Sebastian A *et al* 2017 Temporal correlation detection using computational phase-change memory *Nat. Commun.* **8** 1115

[118] Gao B, Wu H, Kang J, Yu H and Qian H 2016 Oxide-based analog synapse: physical modeling, experimental characterization, and optimization *IEEE Int. Electron Devices Meeting* pp 7.3.1–4

[119] Gao B *et al* 2017 Modeling disorder effect of the oxygen vacancy distribution in filamentary analog RRAM for neuromorphic computing *IEEE Int. Electron Devices Meeting* pp 4.4.1–4

[120] Park S *et al* 2013 Neuromorphic speech systems using advanced ReRAM-based synapse *IEEE Int. Electron Devices Meeting* pp 25.6.1–4

[121] Peng C *et al* 2012 W–Sb–Te phase-change material: a candidate for the trade-off between programming speed and data retention *Appl. Phys. Lett.* **101** 122108

[122] Loke D *et al* 2012 Breaking the speed limits of phase-change memory *Science* **336** 1566–9

[123] Muraoka S, Ninomiya T, Wei Z, Katayama K, Yasuhara R and Takagi T 2013 Comprehensive understanding of conductive filament characteristics and retention properties for highly reliable ReRAM *Symp. on VLSI Technology* pp T62–3

[124] Suri M *et al* 2012 CBRAM devices as binary synapses for low-power stochastic neuromorphic systems: auditory (cochlea) and visual (retina) cognitive processing applications *Int. Electron Devices Meeting* pp 10.3.1–4

[125] Ninomiya T *et al* 2012 Conductive filament scaling of TaO_x bipolar ReRAM for long retention with low current operation *Symposium on VLSI Technology (VLSIT)* pp 73–4

[126] Wu H *et al* 2017 Resistive random access memory for future information processing system *Proc. IEEE* **105** 1770–89

[127] Chen Y Y *et al* 2013 Endurance/retention trade-off on HfO_2/metal cap 1T1R bipolar RRAM *IEEE Trans. Electron Devices* **60** 1114–21

[128] Wu E *et al* 2017 Fundamental limitations of existing models and future solutions for dielectric reliability and RRAM applications *IEEE Int. Electron Devices Meeting* pp 21.5.1–4

[129] Gao B *et al* 2011 Modeling of retention failure behavior in bipolar oxide-based resistive switching memory *IEEE Electron Device Lett.* **32** 276–8

[130] Chen H *et al* 2014 Towards high-speed, write-disturb tolerant 3D vertical RRAM arrays *Symp. on VLSI Technology: Digest of Technical Papers* pp 1–2

[131] Ielmini D, Nardi F and Cagli C 2010 Resistance-dependent amplitude of random telegraph-signal noise in resistive switching memories *Appl. Phys. Lett.* **96** 53503

[132] Kang J *et al* 2017 Time-dependent variability in RRAM-based analog neuromorphic system for pattern recognition *IEEE Int. Electron Devices Meeting* pp 6.4.1–4

[133] Chen P, Peng X and Yu S 2017 NeuroSim+: an integrated device-to-algorithm framework for benchmarking synaptic devices and array architectures *IEEE Int. Electron Devices Meeting* pp 6.1.1–4

[134] Lecun Y, Bottou L, Bengio Y and Haffner P 1998 Gradient-based learning applied to document recognition *Proc. IEEE* **86** 2278–324

[135] Gao L *et al* 2015 Fully parallel write/read in resistive synaptic array for accelerating on-chip learning *Nanotechnology* **26** 455204

[136] Gao L, Chen P and Yu S 2015 Programming protocol optimization for analog weight tuning in resistive memories *IEEE Electron Device Lett.* **36** 1157–9

[137] Zhang W *et al* 2015 An electronic synapse device based on solid electrolyte resistive random access memory *IEEE Electron Device Lett.* **36** 772–4

[138] Wang Z *et al* 2016 Engineering incremental resistive switching in TaO_x based memristors for brain-inspired computing *Nanoscale* **8** 14015–22

[139] Yu S, Wu Y, Jeyasingh R, Kuzum D and Wong H S P 2011 An electronic synapse device based on metal oxide resistive switching memory for neuromorphic computation *IEEE Trans. Electron Devices* **58** 2729–37

[140] Yuan F-Y *et al* 2017 Conduction mechanism and improved endurance in HfO_2-based RRAM with nitridation treatment *Nanoscale Res. Lett.* **12** 574

[141] Govoreanu B *et al* 2015 A-VMCO: a novel forming-free, self-rectifying, analog memory cell with low-current operation, nonfilamentary switching and excellent variability *2015 Symp. on VLSI Technology* pp T132–3

[142] Bichler O, Suri M, Querlioz D, Vuillaume D, DeSalvo B and Gamrat C 2012 Visual pattern extraction using energy-efficient '2-PCM synapse' neuromorphic architecture *IEEE Trans. Electron Devices* **59** 2206–14

[143] Djurfeldt M, Lundqvist M, Johansson C, Rehn M, Ekeberg O and Lansner A 2008 Brain-scale simulation of the neocortex on the IBM Blue Gene/L supercomputer *IBM J. Res. Dev.* **52** 31–41

[144] Eryilmaz S B *et al* 2016 Training a probabilistic graphical model with resistive switching electronic synapses *IEEE Trans. Electron Devices* **63** 5004–11

[145] George D and Hawkins J 2009 Towards a mathematical theory of cortical micro-circuits *PLoS Comput. Biol.* **5** 1–26

[146] Schemmel J, Fieres J and Meier K 2008 Wafer-scale integration of analog neural networks *IEEE Int. Joint Conf. on Neural Networks (IEEE World Congress on Computational Intelligence)* pp 431–8

[147] Benjamin B V *et al* 2014 Neurogrid: a mixed-analog-digital multichip system for large-scale neural simulations *Proc. IEEE* **102** 699–716

[148] Chen P-Y *et al* 2015 Technology-design co-optimization of resistive cross-point array for accelerating learning algorithms on chip *Design, Automation Test in Europe Conf. Exhibition* pp 854–9

[149] Raina R, Madhavan A and Ng A Y 2009 Large-scale deep unsupervised learning using graphics processors *Proc. 26th Annual Int. Conf. on Machine Learning* pp 873–80

[150] Furber S B, Galluppi F, Temple S and Plana L A 2014 The SpiNNaker project *Proc. IEEE* **102** 652–65

[151] Akopyan F *et al* 2015 TrueNorth: design and tool flow of a 65 mW 1 million neuron programmable neurosynaptic chip *IEEE Trans. Comput. Des. Integr. Circuits Syst.* **34** 1537–57

[152] Le Q V *et al* 2012 Building high-level features using large scale unsupervised learning *Proc. 29th Int. Conf. on Int. Conf. on Machine Learning* pp 507–14

[153] Yu S, Lee B and Wong H S P 2012 Metal oxide resistive switching memory *Functional Metal Oxide Nanostructures* ed J Wu *et al* (New York: Springer) pp 303–35

[154] (https://newsroom.intel.com/news-releases/intel-and-micron-produce-breakthrough-memory-technology/)

[155] Chang C *et al* 2018 Mitigating asymmetric nonlinear weight update effects in hardware neural network based on analog resistive synapse *IEEE J. Emerg. Sel. Top. Circuits Syst.* **8** 116–24

[156] Yang C-S, Shang D-S, Chai Y-S, Yan L-Q, Shen B-G and Sun Y 2017 Electrochemical-reaction-induced synaptic plasticity in MoO_x-based solid state electrochemical cells *Phys. Chem. Chem. Phys.* **19** 4190–8

[157] Jang J, Park S, Burr G W, Hwang H and Jeong Y 2015 Optimization of conductance change in $Pr_{1-x}Ca_xMnO_3$-based synaptic devices for neuromorphic systems *IEEE Electron Device Lett.* **36** 457–9

[158] Kim S, Kim H, Hwang S, Kim M-H, Chang Y-F and Park B-G 2017 Analog synaptic behavior of a silicon nitride memristor *ACS Appl. Mater. Interfaces* **9** 40420–7

[159] Woo J, Padovani A, Moon K, Kwak M, Larcher L and Hwang H 2017 Linking conductive filament properties and evolution to synaptic behavior of RRAM devices for neuromorphic applications *IEEE Electron Device Lett.* **38** 1220–3