Roadmap to neuromorphic computing with emerging technologies

Cite as: APL Mater. 12, 109201 (2024); doi: 10.1063/5.0179424
Submitted: 2 October 2023 • Accepted: 20 August 2024 •
Published Online: 21 October 2024

Adnan Mehonic,[1,a)] Daniele Ielmini,[2] Kaushik Roy,[3] Onur Mutlu,[4] Shahar Kvatinsky,[5] Teresa Serrano-Gotarredona,[6] Bernabe Linares-Barranco,[6] Sabina Spiga,[7] Sergey Savel'ev,[8] Alexander G. Balanov,[8] Nitin Chawla,[9] Giuseppe Desoli,[10] Gerardo Malavena,[2] Christian Monzio Compagnoni,[2] Zhongrui Wang,[11] J. Joshua Yang,[11] Syed Ghazi Sarwat,[12] Abu Sebastian,[12] Thomas Mikolajick,[13,14] Stefan Slesazeck,[13] Beatriz Noheda,[15] Bernard Dieny,[16] Tuo-Hung (Alex) Hou,[17] Akhil Varri,[18] Frank Brückerhoff-Plückelmann,[18] Wolfram Pernice,[18] Xixiang Zhang,[19] Sebastian Pazos,[19] Mario Lanza,[20] Stefan Wiefels,[21] Regina Dittmann,[21] Wing H. Ng,[1] Mark Buckwell,[1] Horatio R. J. Cox,[1] Daniel J. Mannion,[22] Anthony J. Kenyon,[1] Yingming Lu,[23] Yuchao Yang,[23] Damien Querlioz,[24] Louis Hutin,[25] Elisa Vianello,[25] Sayeed Shafayet Chowdhury,[3] Piergiulio Mannocci,[2] Yimao Cai,[26] Zhong Sun,[26] Giacomo Pedretti,[27] John Paul Strachan,[28,29] Dmitri Strukov,[30] Manuel Le Gallo,[12] Stefano Ambrogio,[31] Ilia Valov,[32,33] and Rainer Waser[32,34]

For affiliations, please see the end of the Reference section

TABLE OF CONTENTS

I. INTRODUCTION TO THE ROADMAP

Adnan Mehonic, Daniele Ielmini, and Kaushik Roy

A. Taxonomy and motivation

The growing adoption of data-driven applications, such as artificial intelligence (AI), is transforming the way we interact with technology. Currently, the deployment of AI and machine learning tools in previously uncharted domains generates considerable enthusiasm for further research, development, and utilization. These innovative applications often provide effective solutions to complex, longstanding challenges that have remained unresolved for years. By expanding the reach of AI and machine learning, we unlock new possibilities and facilitate advancements in various sectors. These include, but are not limited to, scientific research, education, transportation, smart city planning, eHealth, and the metaverse.

However, our predominant focus on performance can sometimes lead to critical oversights. For instance, our constant dependence on immediate access to information might cause us to ignore the energy consumption and environmental consequences associated with the computing systems that enable such access. Balancing performance with sustainability is crucial for the technology's continued growth.

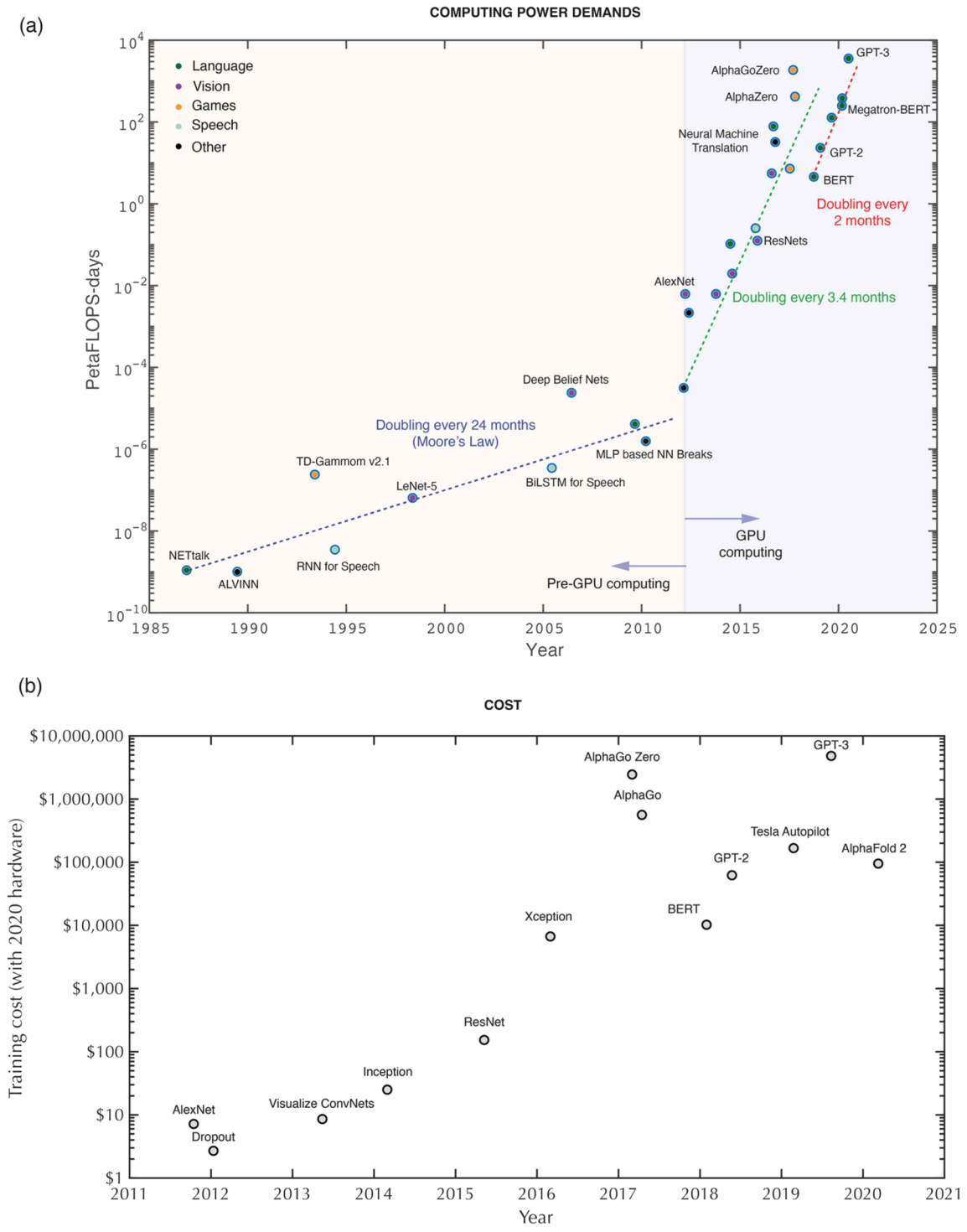

FIG. 1. (a) Increase in computing power demands to run state-of-the-art AI models. (b) The cost associated with training AI models. Adapted and reproduced with permission from A. Mehonic and A. J. Kenyon, Nature **604**, 255 (2022). Copyright 2022 Springer Nature Limited.

From this standpoint, the environmental impact of AI is a cause for growing concern. In addition, applications such as the Internet of Things (IoT) and autonomous robotic agents may not always rely on resource-intensive deep learning algorithms but still need to minimize energy consumption. Realizing the vision of IoT is contingent upon reducing the energy requirements of numerous connected devices. The demand for computing power is growing at a rate that far exceeds improvements achieved through Moore's law scaling. Figure 1(a) shows the computing power demands, quantified in peta-floating-point operations (petaflops, one peta = 10^{15}) per day, as a function of time, indicating an increase of a factor 2 every two months in recent years.[1] In addition to Moore's law, significant advancements have been made through the combination of intelligent architecture and hardware–software co-design. For instance, NVIDIA graphics processing units (GPUs') performance has improved by a factor of 317 from 2012 to 2021, surpassing expectations based on Moore's law alone. Research and development efforts have demonstrated further impressive performance improvements,[2–4] suggesting that more can be achieved. However, conventional computing solutions alone are unlikely to meet the demand in the long term, particularly when considering the high costs of training associated with the most complex deep learning models [Fig. 1(b)]. It is essential to explore alternative approaches to tackle these challenges and ensure the long-term sustainability of AI's rapid advancements. While global energy consumption is crucial and important, there is a relevant issue that is perhaps just as significant: the ability of low-power systems to execute complex AI algorithms without relying on cloud-based computing. It is important to keep in mind that the challenge of global AI power consumption and the ability to implement complex AI on low-power systems are two somewhat separate challenges. It might be the case that these two challenges need to be addressed with somewhat different strategies (e.g., the power consumption in data centers for the most complex, largest AI models, such as large language models, might be addressed differently than implementing mid-sized AI models, such as voice recognition, on low-power, self-contained systems that might need to run at a few milliwatts of power). The latter strategy might not be scalable for the largest models, or the optimization of the largest models might not be applicable for simpler models running on much lower power budgets. However, undeniably, for both, we need to improve the overall energy efficiency of our computing systems that are designed to execute AI workloads.

The energy efficiency and performance of computing can largely benefit from new paradigms that aim at replicating or being inspired by specific characteristics of the brain's biological mechanisms. It is important to note that biological systems might be highly specialized and heterogeneous, and therefore, different tasks are addressed by different computational schemes. However, we can still aim to take inspiration from general features when they are advantageous for specific applications. It is unlikely that a single architecture or broader approach will be best applicable for all targeted applications.

Adopting an interdisciplinary methodology, experts in materials science, device and circuit engineering, system design, and algorithm and software development are brought together to collectively contribute to the progressive field of neuromorphic engineering and computing. This collaborative approach is instrumental in fueling innovation and promoting advancements in a domain that seeks to bridge the gap between biological systems and artificial intelligence. Coined by Mead in the late 1980s,[5] the term "neuromorphic" originally referred to systems and devices replicating certain aspects of biological neural systems, but now, it varies across different research communities. While the term's meaning continues to evolve, it generally refers to a system embodying brain-inspired properties, such as in-memory computing, hardware learning, spike-based processing, fine-grained parallelism, and reduced precision computing. One can also draw analogies and identify more complex phenomenological similarities between biological units (e.g., neurons) and electronic components (e.g., memristors). For example, phenomenological similarities between models of the redox-based nanoionic resistive memory cell and common neuronal models, such as the Hodgkin–Huxley conductance model and the leaky integrate-and-fire model, have been demonstrated.[6] Even more complex biological functionalities have been demonstrated using a single third-order nanocircuit elements.[7] It should be noted that many paradigms related to the neuromorphic approach have also been independently investigated. For instance, in-memory computing,[8] while being a cornerstone of the neuromorphic paradigm, is also examined separately. It represents one of the most promising avenues to enhance the energy efficiency of AI hardware or more general computing, offering a break from the traditional von Neumann architecture paradigm.

Neuromorphic research can be divided into three areas. First, "neuromorphic engineering" employs either complementary metal–oxide–semiconductor (CMOS) technology (e.g., transistors working in a sub-threshold regime) or cutting-edge post-CMOS technologies to reproduce the brain's computational units and mechanisms. Second, "neuromorphic computing" explores new data processing methods, frequently drawing inspiration from biological systems and considering alternative algorithms, such as spike-based computing. Finally, the development of "neuromorphic devices" marks the third field. Taking advantage of advancements in electronic and photonic technologies, it develops innovative nanodevices that frequently emulate biological components, such as neurons and synapses, or efficiently implement desired properties, such as in-memory computing.

Furthermore, various approaches to neuromorphic research can be identified based on their primary objectives. Some systems focus on delivering efficient hardware platforms to enhance our understanding of biological nervous systems, while others employ brain-inspired principles to create innovative, efficient computing applications. This roadmap primarily focuses on the latter. While there are already outstanding roadmaps,[9] reviews,[10–12] and special issues[13] that offer comprehensive overviews of neuromorphic technologies, encompassing the integration of hardware and software solutions as well as the exploration of new learning paradigms, this particular roadmap focuses on emphasizing the significance of materials engineering in advancing cutting-edge complementary metal–oxide–semiconductor (CMOS) and post-CMOS technologies. Simultaneously, it offers a holistic perspective on the general challenges of computing systems, the reasoning behind adopting the neuromorphic approach, and concise summaries of current technologies to better contextualize the role of materials engineering within the broader neuromorphic landscape. Of course, there are other critical aspects in the development of neuromorphic tech-

nologies that need to be taken into account. For example, an excellent recent review on thermal management materials, devices, and networks is one such example.[14]

This roadmap is organized into several thematic sections, outlining current computing challenges, discussing the neuromorphic computing approach, analyzing mature and currently utilized technologies, providing an overview of emerging technologies, addressing material challenges, exploring novel computing concepts, and finally examining the maturity level of emerging technologies while determining the next essential steps for their advancement.

This roadmap starts with a concise introduction to the current digital computing landscape, primarily characterized by Moore's law scaling and the von Neumann architecture. It then explores the challenges in sustaining Moore's law and examines the significance and potential advantages of post-CMOS technologies and architectures aiming to integrate computing and memory. Following this, this roadmap presents a historical perspective on the neuromorphic approach, emphasizing its potential benefits and applications. It provides a thorough review of cutting-edge developments in various emerging technologies, comparing them critically. The discussion addresses how these technologies can be utilized to develop computational building blocks for future computing systems. The roles of two mature technologies, static random access memory (SRAM) and flash, are also explored. The overview of emerging technologies includes resistive switching and memristors, phase change materials, ferroelectric materials, magnetic materials, spintronic materials, optoelectronic and photonic materials, and 2D devices and systems. Material challenges are discussed in detail, covering types of challenges, possible solutions, and experimental techniques to study these. Novel computing concepts are examined, focusing on embracing device and system variability, spiking-based computing systems, analog computing for linear algebra, and the use of analog content addressable memory (CAM) for in-memory computing and optimization solvers. Section VIII discusses technological maturity and potential future directions.

II. COMPUTING CHALLENGES

Onur Mutlu and Shahar Kvatinsky

A. Digital computing

1. Status

Digital computing has a long and complex history that stretches back over a century. The earliest electronic computers were developed in the 1930s and 1940s, and they were large, expensive, and difficult to use. However, these early computers laid the foundation for the development of the modern computers that we use today and their principles are still in widespread use.

One of the key figures in the early history of digital computing was John von Neumann, a mathematician and computer scientist known for his contributions to the field of computer science. Von Neumann advocated the stored program concept and sequential instruction processing, two vital features of the von Neumann architecture[15] that are still used in most computers today. Another key feature of the von Neumann architecture is the separation of the central processing unit (CPU) and the main memory. This separation allows the CPU to access the instructions and data it needs from the main memory while executing a program, and assigns the computation and control responsibilities specifically to the CPU.

Throughout the years, the rapid scaling of semiconductor logic technology, known as Moore's law,[16] has led to tremendous improvements in computer performance and energy efficiency. With the exponential increase in the number of transistors placed on a single chip provided by technology scaling, engineers have explored many ways to increase the speed and performance of computers. One way they did this was by exploiting parallelism, which is the ability of a computer to perform multiple tasks simultaneously. There are several different types of parallelism, including SISD (single instruction, single data), SIMD (single instruction, multiple data), MIMD (multiple instruction, multiple data), and MISD (multiple instruction, single data),[17] all of which are exploited in modern computing systems ranging from general-purpose single-core and multi-core processors, GPUs, and specialized accelerators.

Technology scaling has also allowed for the development of more processing units, starting from duplicating the processing cores and, more recently, adding accelerators. These accelerators can off-load specific tasks, e.g., video processing, compression/decompression, vision processing, graphics, and machine learning, from the central processor, further improving performance and energy efficiency (the required energy to perform a certain task) by specializing the computation units to the task at hand. As such, modern systems are heterogeneous, with many different types of logic-based computation units integrated into the same processor die.

2. Challenges

While the performance and energy of logic-based computation units have scaled very well via technology scaling, those of interconnect and memory systems have not scaled as well. As a result, communication (e.g., data movement) between computation units and memory units has emerged as a major bottleneck, partly due to the disparity in scaling and partly due to the separation and disparity between processing and memory offered in von Neumann architecture, which both have limited the ability of computers to take full advantage of the improvements in logic technology. This bottleneck is broadly referred to as the "von Neumann bottleneck" or the "memory wall," as it can greatly limit the speed and energy at which the computer can execute instructions.

For decades, the transistor size has scaled down, while the power density has remained constant. This phenomenon, first observed in the 1970s by Dennard,[18] means that as transistors become smaller and more densely packed onto a chip, the overall performance and capabilities of the chip improve. However, since the early 2000s, it has become increasingly challenging to maintain Dennard scaling as voltage (and thus frequency) scaling has greatly slowed down. The end of Dennard scaling has increased the importance of energy efficiency of different processing units and led to phenomena such as "dark silicon,"[19] where large parts of the chip are powered off. The rapid move toward more specialized processing units, powered on for specific tasks, exemplifies the influence of the end of Dennard scaling.

Furthermore, in recent years, it has become increasingly challenging to maintain the pace of Moore's law due to the physical lim-

itations of transistors and the challenges of manufacturing smaller and more densely packed chips. As a result, the looming end of Moore's law has been a topic of discussion in the tech industry, as this could potentially limit the future performance improvements of computer chips. New semiconductor technologies and novel architectural solutions are required to continue computing systems' performance and energy efficiency improvements at a similar pace as in the past.

3. Potential solutions and conclusion

In recent years, different semiconductor and manufacturing technologies have emerged to overcome the slowdown of Moore's law. These devices include new transistor structures and materials, advanced packaging techniques, and new (e.g., non-volatile) memory devices. Some of those technologies have similar functionality as standard CMOS technology but with improved properties. Other technologies also offer radically new properties, different from CMOS. For example, memristive technologies, such as resistive RAM,[20] have varying resistance and provide analog data storage that also supports computation. Such novel technologies with their unique properties may serve as enablers for new architectures and computing paradigms, which could be different from and complementary to the von Neumann architecture.

The combination of Moore's law slowdown and von Neumann's bottleneck requires fresh thinking on computing paradigms. Data movement between the memory and the processing units is the primary impediment against high performance and high energy efficiency in modern computing systems.[21–24] In addition, this impediment only worsens with the improved processing abilities and the increased need for data. All modern computers employ a variety of methods to mitigate the memory bottleneck, all of which increase the complexity and power requirements of the system with limited (and sometimes little) success in mitigating the bottleneck. For example, modern computers have several levels of cache memories to reduce the latency and power of memory accesses by exploiting data locality. Cache memories, however, have limited capacity and are effective only when significant spatial and temporal locality exists in the program. Cache memories are not always (completely) effective due to low locality in many modern workloads, which can worsen the performance and energy efficiency of computers.[25,26] Similarly, modern computers employ prefetching techniques across the memory hierarchy to anticipate future memory accesses and load data into caches before they are needed by the processor. While partially effective for relatively simple memory access patterns, prefetching is not effective for complicated memory access patterns and it increases system complexity and memory bandwidth consumption.[27] Thus, memory bottleneck remains a tough challenge and hundreds of research papers and patents are written every year to mitigate it.[28]

Overcoming the performance and energy costs of off-chip memory accesses is an increasingly difficult task as the disparity between the efficiency of computation and the efficiency of memory access continues to grow. There is therefore a need to examine more disruptive technologies and architectures that much more tightly integrate logic and memory at a large scale, avoiding the large costs of data movement across system components.

Many efforts to move computation closer to and inside the memory units have been made,[29] including adding processing units in the same package as DRAM chips,[30,31] performing digital processing using memory cells,[32,33] and using analog computation capabilities of both DRAM and non-volatile memory (NVM) devices.[34–37] One exciting novel computing paradigm to eliminate the von Neumann bottleneck is to reconsider the way computation and memory tasks are performed by getting inspiration from the brain, where, unlike von Neumann architecture, processing and storage are not separated. Many recent studies demonstrate orders of magnitude performance and energy improvements using various kinds of processing-in-memory architectures.[29] Processing-in-memory and, more broadly, neuromorphic (or brain-inspired) computing thus offers a promising way to overcome the major performance and energy bottleneck in modern memory systems. However, it also introduces significant challenges for adoption as it is a disruptive technology that affects all levels of the system stack, from hardware devices to software algorithms.

III. NEUROMORPHIC COMPUTING BASICS AND ITS EVOLUTION

Teresa Serrano-Gotarredona and Bernabe Linares Barranco

A. What is neuromorphic computing/engineering

Neuromorphic computing can be defined as the underlying computations performed by neuromorphic physical systems. Neuromorphic physical systems carry out robust and efficient neural computation using hardware implementations that operate in physical time. Typically, they are event- or data-driven and they employ low-power, massively parallel hybrid analog, digital, or mixed VLSI circuits, and they operate using similar physics of computation used by the nervous system.

Spiking neural networks (SNNs) are one very good example of a neuromorphic computing system. Computation is performed whenever a spike is transmitted and received by destination neurons. Computation can be performed at the dendritic tree, while spikes travel to their destinations, as well as at the destination neurons where they are collected to update the internal states of the neurons. Neurons collect pre-weighted and pre-filtered spikes coming from different source neurons or sensors, perform some basic computation on them, and generate an output spike whenever their internal state reaches some threshold. A neuron firing typically means that the "feature" this neuron represents has been identified in place and time. The collective computation of populations of neurons can give rise to powerful system level behaviors, such as pattern recognition, decision making, sensory fusion, and knowledge abstraction. In addition, neuromorphic computing systems can also be enabled to acquire new knowledge through both supervised and unsupervised learning, either offline or while they perform, which is typically known as online learning and which can be life-long. Neuromorphic computing covers typically from sensing to processing to learning.

1. Neuromorphic sensing

Probably the most clarifying example of what neuromorphic computation is about is the paradigm of neuromorphic visual computation. Neuromorphic visual computation exploits the data

encoding provided by neuromorphic visual sensors. At present, the most widespread neuromorphic vision sensor is the Dynamic Vision Sensor (DVS).[38] In a DVS, each pixel sends out its (x, y) coordinate whenever its photodiode perceives a relative change of light beyond some preset thresholds $\theta^- > \frac{I_{n+1}}{I_n} > \theta^+$, with θ^+ slightly greater than 1 and θ^- slightly less than 1. This is typically referred to as an "address event." If $I_{n+1} > I_n$, then light has increased. If $I_{n+1} < I_n$, then light has decreased. To differentiate both situations, the address event can also be a signed event, by adding a sign bit "s," (x, y, s). If events are recorded using some event-recording hardware, then a timestamp t_n is added to each event (x_n, y_n, s_n, t_n). The full recording consists then of a list of timestamped address events. Figure 2 illustrates this. In Fig. 2(a), a DVS camera is observing a 7 kHz spiral on a classic phosphor oscilloscope (without any extra illumination source). Figure 2(b) plots in $\{x, y, t\}$ space the recorded events. The camera was a 128 × 128 pixel high-contrast sensitivity DVS camera.[39] Therefore, x–y coordinates in Fig. 2(b) span from 0 to 127. The vertical axis is time, which spans over about 400 μs, slightly less than 100 μs per spiral turn. Each dot in Fig. 2(b) is an address event, and we can count several hundreds of them within the 400 μs. This DVS camera is capable of generating over 10×10^6 events per second (about one every 100 ns). This produces a very fine timing resolution when sensing dynamic scenes.

The information (events) produced by this type of sensors can be sent directly to event-driven neuromorphic computing hardware, which would process this quasi-instantaneous dynamic visual information event by event.

DVS cameras have evolved over the past 20 years, since they first appeared.[40] They combine frames, sensitivity to color,[41] and of resolutions up to 1 MP.[42]

Other sensory modality event-driven neuromorphic devices have been reported, such as auditory cochleae,[43] olfactory noses,[44] or tactile sensing.[45]

2. Neuromorphic processing

Neuromorphic signal information encoding in the form of sequences of events reduces information so that only meaningful data, such as changes, are transmitted and processed. This follows the underlying principle in biological nervous systems, as information transmission (in the form of nervous spikes) and their consequent processing affect energy consumption. Thus, biological systems tend to minimize the number of spikes (events) to be transmitted and processed for a given computational task. This principle is what neuromorphic computing intends to pursue. Figure 3 shows an illustrative example of this efficient frame-free event-driven information encoding.[46] In Fig. 3(a), we see a poker card deck being browsed at a natural speed, recorded with a DVS, and played back at real-time speed with a reconstructed frame time of about 20 ms. In Fig. 3(b), the same recorded list of events is played back at 77 μs frame time. In Fig. 3(c), we show the tracked symbol input fed to a spiking convolutional neural network for object recognition, displaying the recognized output symbol. In Fig. 3(d), we show the 4-layer spiking convent structure, and in Fig. 3(e), we show the $\{x, y,$ time$\}$ representation of 20 ms input and output events occurring during a change of card so that the recognition switches from one symbol to the next in less than 2 ms. Note that here the system is composed of both, the sensor and the network executing the recognition. Both working together need less than 2 ms. This contrasts dramatically with conventional artificial systems, in which the sensor first needs to acquire two consecutive images (typically 25 ms per image) and then process both to capture the change.

Figure 3 illustrates a simple version of a neuromorphic sensing and processing system. By today, much larger neuromorphic systems, inspired in the same information encoding scheme, have been developed and demonstrated. The following are some powerful example systems:

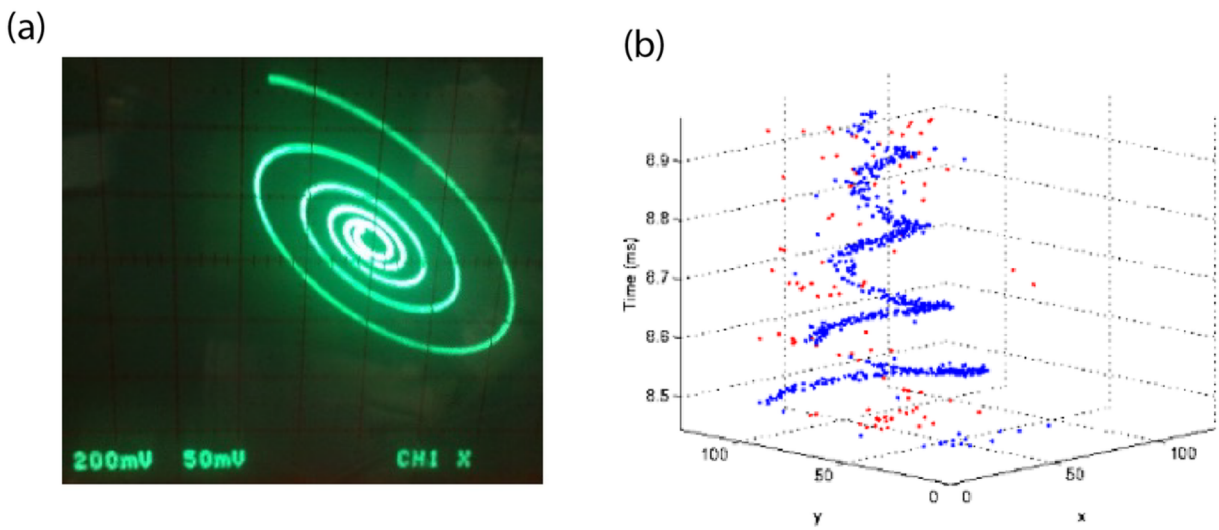

FIG. 2. (a) 7 kHz spiral observed in a classic phosphor oscilloscope set in X/Y mode. (b) DVS output event stream when observing the oscilloscope in (a).

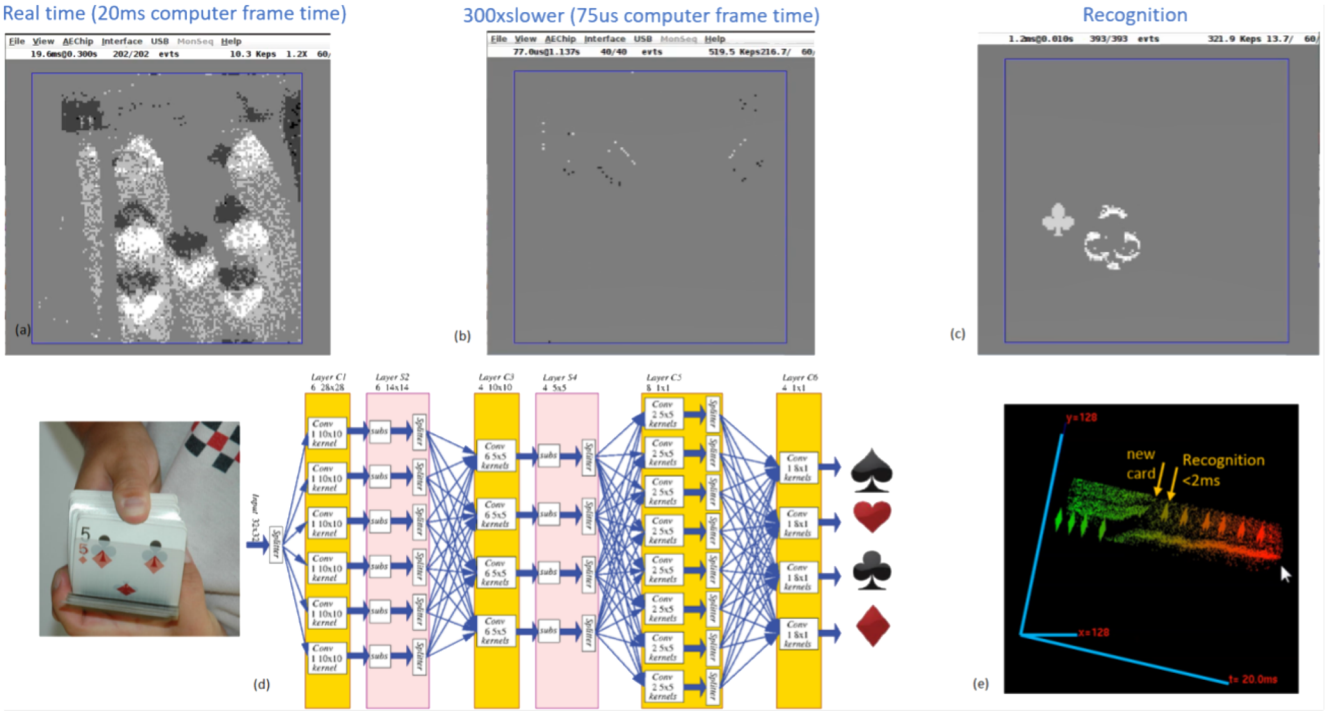

FIG. 3. (a) Fast speed poker deck browsing: events are collected about every 20 ms to build a frame to display on a computer screen. (b) Slow speed playback at 77 μs per reconstructed frame. (c) Poker symbol tracked and displayed on the right and recognition output on the left. (d) Event-driven CNN to classify four poker symbols. (e) $\{x, y, t\}$ representation of 20 ms showing camera events together with recognition events during a change of card with a recognition switching of less than 2 ms.

- The SpiNNaker platform[47] was partly developed under the Human Brain Project.[48] It features an 18-core Advanced RISC Machine (ARM) SpiNNaker chip.Each node on the platform comprises a printed circuit board (PCB) that holds 48 of these chips. In total, about 1200 chips are assembled into furniture-like sets that include racks. Collectively, these setups host approximately 1 million ARM cores. This system is capable of emulating 1×10^9 neurons in real time. An updated SpiNNaker chip has already been developed, performing about 10× in efficiency, neuron emulation capability, and event traffic handling, while keeping similar power consumption.
- The BrainScales platform,[49] also developed during the Human Brain Project,[47] implements physical silicon neurons fabricated on full silicon 8 in. wafers, interconnecting 20 of these wafers in a cabinet, together with 48 FPGA based communication modules. It implements accelerated time computations with respect to real time (about 10 000×), with spike-timing-dependent plastic synapses. Each wafer can host about 200k neurons and 44×10^6 synapses.
- The IBM TrueNorth chip[50] could host 1×10^6 very simple neurons, or be reconfigured to achieve a trade-off between the number of neurons and neuron model complexity. They were structured into 4096 identical cores, consuming about 63 mW each.
- Loihi from Intel is probably by today the most advanced neuromorphic chip. In its first version,[51] fabricated in 14 nm, it contains 128 cores, each capable of implementing 1k spiking neuronal units (compartments), and includes plastic synapses. More recently, Loihi 2 chip was introduced, with up to 1×10^6 neurons per chip, manufactured in Intel 4 technology (7 nm). Up to 768 of Loihi chips have been assembled into the Pohoiki Springs system, while operating at less than 500 W.[52]

3. Challenges and conclusion

Neuromorphic computing algorithms should be optimum when run on neuromorphic hardware, where events travel and are processed in a fully parallel manner. One of the main challenges in present day neuromorphic computing is to train and execute powerful computing systems directly on neuromorphic hardware. Traditionally, neuromorphic computing problems were mapped to more traditional deep neural networks to obtain their parameters through backpropagation based training,[53] which would then be mapped to their neuromorphic/spiking counterpart.[46] However, these transformations always resulted in a loss of performance. By today, there are many proposals of training directly in the spiking domain, combining variants of spike-timing-dependent plasticity rules, with surrogate training techniques that adapt backpropagation to spiking systems, which are tested on either fully connected or convolution based deep spiking neural networks. For an updated review, readers are referred to Ref. 54.

On the other hand, it remains to see whether novel nanomaterial devices, such as memristors, can provide truly giga-scale compact chips with billions of neurons on a single chip and self-learning algorithms. Some initial demonstrations of single[55] or multi-core systems[56] exploiting a nano-scale memristor combined with a selector transistor as a synaptic element have been reported, with highly promising outlooks once synapse elements could be provided as pure nanoscale devices while stacking multiple layers of synapse fabrics together with other nano-scale neurons.[57] In the end, the success of neuromorphic computing will rely on the optimum combined progress in neuromorphic hardware, most probably exploiting emerging nano-scale devices massively, in event- and data-driven information and energy-efficient processing methodologies, and finally in providing efficient, resilient, and quick learning methodologies for mapping real-world applications into the available hardware and computational neuromorphic substrates.

B. Different neuromorphic technologies and state of the art

Sabina Spiga

1. Status

The research field of neuromorphic computing has been growing significantly over the past three decades, following the pioneering research at Caltech (USA) by Mead,[5] and it is currently attracting the interest of a wide and interdisciplinary community from devices, circuits, and systems to neuroscience, biology, computer science, materials, and physics. Within this framework, the developed neuromorphic *hardware technologies* span from fully CMOS-based systems[58,59] to solutions exploiting the use of charge-based or resistive non-volatile memory technologies[60–62] and to emerging memristive device concepts and novel materials.[63–66] Figure 4 reports a schematic (and non-exhaustive) evolution of the main technologies of interest. A common feature of these approaches is to take inspiration from the brain computation, by co-locating memory and processing [in-memory computing (IMC) approach], to overcome the von Neumann bottleneck. Hardware artificial neural networks (ANNs) can implement IMC computing and provides an efficient physical substrate for machine learning algorithms and artificial intelligence (AI). On the other hand, spiking neural networks (SNNs), encoding and processing information using spikes, hold great promise for applications requiring always-on real-time processing of sensory signals, for example in edge computing, personalized medicine, and Internet of things.

In terms of the maturity of neuromorphic technologies, we can discuss three main blocks.

(i) Current *large-scale hardware neuromorphic computing systems* are fully CMOS-based and exploit digital or analog/mixed-signal technologies. Examples of fabricated chips are the IBM TrueNorth, Intel Loihi, Tianjic, ODIN, and others as discussed in previous review papers.[58,59] In these systems, the neuron and synapse functionalities are emulated by using circuit blocks based on CMOS transistors, capacitors, and volatile SRAM memory. The scientific community is now exploiting these chips to implement novel algorithms for AI applications.

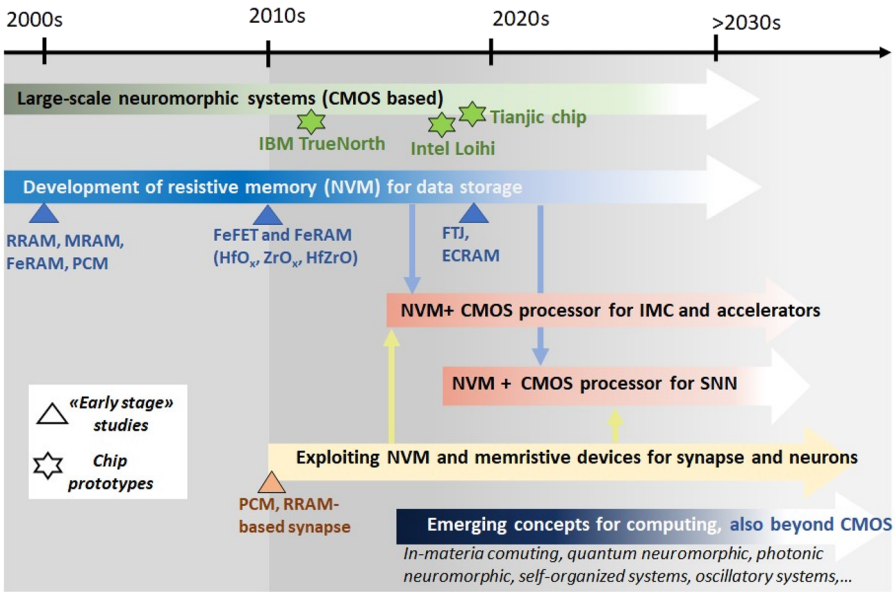

FIG. 4. Schematic evolution of the main hardware technologies of interest for neuromorphic computing (the decades represent only a time frame). Triangular symbols mark the reference period for early stage studies or starting interest in the technology development. From bottom to top of the figure, the listed technologies are today at higher maturity level and more advanced at system integration level.

(ii) *Non-volatile memory technologies*. In the past decade, resistive non-volatile memory (NVM) technologies, such as Resistive Random Access Memory (RRAM), phase change memory (PCM), ferroelectric memory (FeRAM) and ferroelectric transistor (FeFET), and magnetoresistive random access memory (MRAM), have been proposed as possible compact, low power, and dynamical elements to implement in hardware the synaptic nodes, replacing SRAMs, or as a key element of neuronal blocks.[60,61,67] While these NVMs have been developed over the past twenty years mainly for data storage applications, and introduced in the market, they can be considered emerging technologies in the field of neuromorphic computing and their great potential is still not fully exploited. Over the past 10 years, novel concepts for computing, based on hybrid CMOS/non-volatile resistive memory circuits and chips,[56] have been proposed in the literature. In parallel, also more conventional charge-based non-volatile memories, such as flash and NRAM, are currently being investigated for IMC since they are mature technologies. Finally, it is worth mentioning the emerging memory technologies that are attracting increasing interest in the field of IMC and neuromorphic computing, namely the ferroelectric tunnel junction (FTJ)[68] and the 3-terminal electrochemical random access memory (ECRAM).[69]

(iii) *Advanced memristive materials, devices, and novel computation concepts* that are currently investigated include 2D materials, organic materials, perovskites, nanotubes, self-assembled nano-objects and nanowire networks, advanced device concepts in the field of spintronics (domain wall, race-trace memory, and skyrmions), devices based on metal–insulator transition (for instance, VO_2-based devices), and volatile memristors.[65,66,70–72] These technologies are currently proofs of concept at a single device level and circuit blocks connecting a reduced number of devices. The computing system is sometimes demonstrated with a mixed hardware/software approach, where the measured device characteristics are used to simulate large systems. Finally, it is worth mentioning the increasing interest in architectures that can exploit photonic components for computing, toward the building of neuromorphic photonic processors taking advantage of the silicon photonic platforms and co-integration with novel optical memory devices and advanced materials, such as phase-change materials.[73,74]

Figure 5 schematically shows examples of the material systems currently most investigated in various approaches and technologies for neuromorphic computing.

2. Challenges

The current and future challenges can be considered at various levels.

(i) For *large-scale neuromorphic processors*, the progress of CMOS-based technologies and their scaling still provide room to advance the research field. The main challenges are at the architecture and algorithm level. On the other hand,

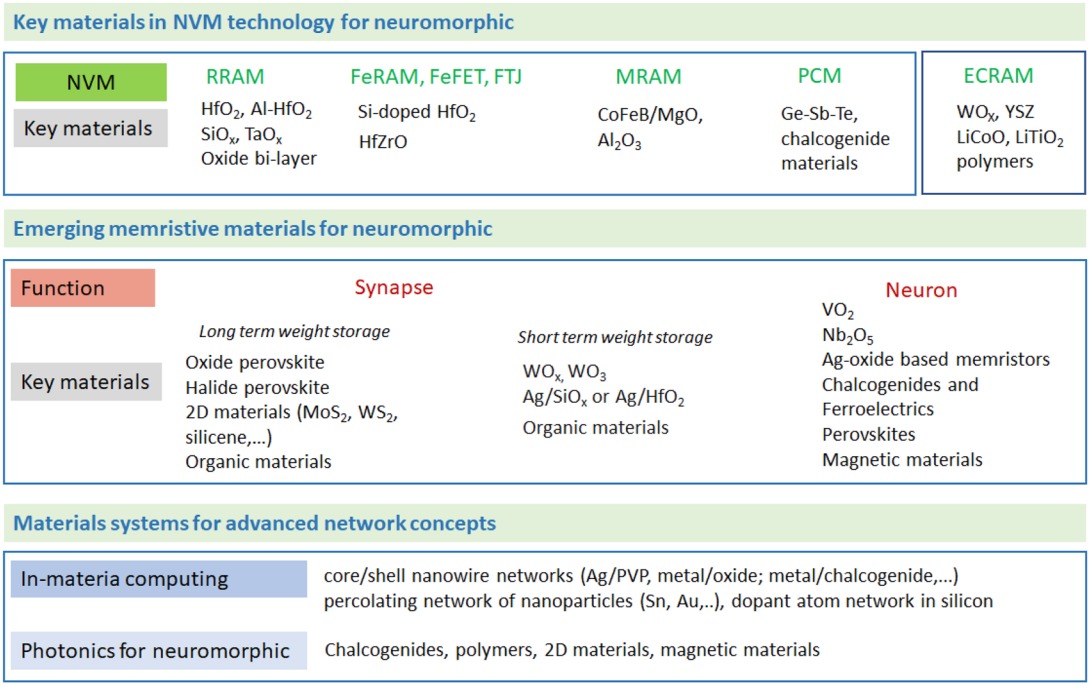

FIG. 5. Examples of materials systems currently employed in memristive technologies. The list of materials is not exhaustive and includes only some of the most used ones. For the NVM devices (top line), the main active material is indicated, but each device includes also various types of material electrodes depending on the technology.

most *NVM memories* (RRAM, PCM, FeRAM, FeFET, and MRAM) have been already integrated with CMOS at scaled technological nodes and large integration density and hold interesting properties (depending on the specific technology), such as small size, scalability, possible easy integration also in 3D array stacking, low programming energy, and multilevel programming capability. Therefore, it is expected that NVM technologies will play an increasing role in future IMC chips or neuromorphic processors, by enabling energy-efficient computation. Prototype IMC chips have been reported in the literature,[56,75] as well as innovative circuits for SNNs implementing advanced learning rules to compute with dynamics.[76,77]

On the other hand, it is worth mentioning that the NVM technologies exhibit several device-level non-idealities, as discussed in more detail in Sec. V of this roadmap. As examples and a non-exhaustive list, we can mention nonlinearity and stochasticity in conductance update vs a number of pulses at a fixed voltage (PCM, RRAM, and FeRAM), asymmetry (RRAM) in the bidirectional tuning of conductance, conductance drift (PCM) or broadening of the resistance distribution (RRAM) after programming, device-to-device and cycle-to-cycle variability of the programmed states, low resolution due to the limited number of programmable levels (up to 8 or 16 are demonstrated for RRAM and PCM at the array level), restricted memory window (MRAM) or limited endurance (general issue except for MRAM), and relatively high conduction also in the OFF state. All these aspects can impact the neural network accuracy and reliability, although proper algorithms/architectures can take advantage from stochasticity or asymmetry of conductance tuning.[78] Therefore, a careful co-design of hardware and algorithms is required together with an improvement of circuit design and/or programming device strategies to fully exploit NVMs in combination with CMOS and in large systems. The specific challenges and possible specific applications of the listed technologies will be discussed further in Secs. V A–V D of this roadmap, while a more deep view on application scenario is reported in Sec. VII.

(ii) Regarding the plethora of *emerging materials/devices and novel concepts* proposed for neuromorphic computing (beyond the ones discussed in the previous point, see some examples in Secs. V D–V F of this roadmap), the main challenge is that they are mostly demonstrated at the single device level or in early stage proofs of concept in small arrays/large device sizes, which are implemented in ANNs or SNNs only at the simulation level. To leverage these concepts at higher technology readiness level (TRL), it is necessary to prove that the device characteristics are reproducible and scalable, the working principle well understood, and to provide more advanced characterizations on several down-scaled devices, and finally to close the current gap between laboratory exploration of single materials/devices and integration in arrays or circuits. Another challenge is to address in more detail how to exploit nanodevices' peculiarities, such as dynamic or stochastic behavior, to implement in hardware more complex bio-inspired functionalities or even to perform radically new computation paradigms. Indeed, while the more standard technologies (CMOS, flash, and SRAM) can also be used in hardware neural networks to implement complex functions, this is possible only at the high cost of increased circuit complexity. To give an example, the required dynamic to reproduce the synaptic or neuronal functionality in SNNs is implemented at the circuit level and/or using large area capacitors, which are not easily scalable in view of large systems. One possible approach is to exploit the emerging memristive technologies and their properties (variability, stochasticity, and non-idealities) to implement complex functions with more compact and low-power devices. One example is the use of resistivity drift in PCM (usually an unwanted characteristic for IMC or storage applications) to implement advanced learning rules in hardware SNNs.[79] Another example (discussed in Sec. VII A) is to use the inherent variability and stochasticity of some nanodevices to build efficient random number generators (for data security applications) and stochastic computing models. Overall, this scenario points out a long-term development research, likely up to ten years or more, to close the gap between these novel concepts and real industrial applications.

3. Potential solutions

To pursue advances in the development of neuromorphic hardware chips, it is necessary to develop a common framework to compare and benchmark different approaches, also in view of some metrics, such as computing density, energy efficiency, computing accuracy, learning algorithms, theoretical framework, and target possible killer applications that might significantly benefit from neuro-inspired chips. Within this framework, materials strategies can still be relevant to address some of the outlined challenges for NVMs, but materials need to be co-developed together with a demonstration of a device at the scaled node and array level. An important strategy for the future is also the possibility to substitute current mainstream materials with green materials or to identify fabrication processes more sustainable in terms of cost and environmental impact, without compromising the hardware functionality. Moreover, other important aspects include the development of hardware architecture that can lead to the integration of several devices and exploiting a large connectivity among them; the implementation of efficient algorithms supporting online learning, also on different time scales as in biological systems; and addressing the low power analysis of large amount of data also for Internet of things and edge devices. Overall, it is a necessary and holistic view that includes the materials/devices/architectures/algorithms co-design to develop a large-scaled neuromorphic chip.

4. Conclusion

The development of advanced neuromorphic hardware that can efficiently support AI applications is becoming more and more important. Despite the several prototypes and results presented in the literature, neuro-inspired chips are still only at an early stage of development and there is plenty of room for further development. Many mature NVM devices are definitely candidates to become a future mainstream technology for large scale neuromorphic processors that can outperform the current platform based only on CMOS circuits. In the long term, it is also necessary to close the gap between emerging materials and concepts, currently demonstrated

only by proofs of concept, and their possible integration in functional systems. Materials research and an understanding of physical principles enabling novel functionalities are important parts for this scenario.

C. Possible future computational primitives for neuromorphic computing

Sergey Savel'ev and Alexander Balanov

The core idea of neuromorphic computing to develop and design computational systems mimicking electrochemical activities in brain cortex is currently booming, embracing areas of deep physical neural networks,[80] classical and quantum reservoir computing,[81,82] oscillator-based computing,[83] and spiking networks,[84] among many other concepts.[85] These computational paradigms imply new ways for information processing and storage different from conventional computing and, therefore, require elementary base and primitives that often involve unusual novel physical principles.[86]

At present, memristors—electronic switchers with memory—and their circuits demonstrate great potential for application in the primitives for future neuromorphic computing systems. In particular, different types of volatile and non-volatile memristors can serve as artificial neurons and synapses, respectively, which facilitate the transfer, storage, and processing of information.[87] For example, volatile Mott memristors[88] can work as an electric oscillator with either regular or chaotic dynamics,[7] while memristors with filament-formation[89] demonstrate tunable stochasticity[90,91] and allow designing neuromorphic circuits with different degrees of plasticity, chaoticity, and stochasticity to address diverse computational aims in mimicking dynamics of different neuron populations. Furthermore, a crossbar of non-volatile memristors (servicing to memorize training) attached to volatile memristors (working as readouts) enables the design of AI hardware with unsupervised learning capability.[92] Thus, combining memristive circuits with different functionalities paves the way for building a wide range of in-memory computational blocks for a broad spectrum of artificial neural networks (ANNs) starting from deep learning accelerators to spiking neuron networks.[93]

A rapidly developing class of volatile memristive elements[94] has been shown to demonstrate a rich spectrum of versatile dynamical patterns,[7,95,96] which makes them suitable for the realization of a range of neuroscience-motivated AI concepts.[97–99] For instance, the ANNs based on volatile memristors can go well beyond usual oscillator-based computing[83] or spiking neural networks.[84] They rely on manipulating information by utilizing complexity in dynamical regimes that offer a novel computational framework[97,98] with cognitive abilities closer to biological brains. There is a specific emphasis on using dynamical behaviors of memristors, instead of only static behaviors.[100]

Remarkably, memristive elements can be realized not only in electronic devices but also within spintronic or photonic frameworks, which have their own advantages compared to electronics. Therefore, hybridized design promises great benefits in the further development of neuromorphic primitives. For example, a combination of memristive chipsets with spintronic and/or photonic components can potentially create AI hardware with enhanced parallelism offered by optical devices operating simultaneously at many frequencies (e.g., optical cavity eigenfrequencies),[101] energy-efficient magnetic non-volatile memories, and flexible memristive spiking network architectures. An important step in the realization of this approach is the development of interface technologies for bringing electronic, photonic, and spintronic technologies together. A possible example is the spintronic memristor,[65,102,103] where the transformation of magnetic structure influences the resistance of the system.[104] An interface between neuromorphic optical and electronic subsystems of a hybrid device could be realized using optically controlled electronic memristive systems,[105] thus paving a path for neuromorphic optoelectronic systems.[106]

The conventional ANNs with a large number of connections require training to be efficient in the task requiring frequent retraining for "moving target" problems, for example in the recognition of characteristics changing in time. A potential solution for such tasks is to implement filtering or pre-processing data by a "reservoir,"[81] usually consisting of neuron units connected by fixed weights. The reservoir is assisted by a small readout ANN, which requires much less data for training, thus removing significant retraining burden. Recently, an important evolution has taken place in the development of reservoir computing systems, where the function of the reservoir is realized by photon, phonon, and/or magnon mode mixing in spintronic[107,108] and photonic[109] devices. Substitution of the interaction of many artificial neurons by wave processes resembles neural wave computation in the visual cortex[98] and promotes miniaturization, robustness, and energy efficiency of the reservoirs (neuromorphic accelerators), which in the future could become an additional class of primitive, especially in neuromorphic computational systems dealing with temporal or sequential data processing.[110] In AI training, it has also been shown that memristive matrix multiplication hardware can enable noisy local learning algorithms, which perform training at the edge with significant energy efficiencies compared to graphics processing units.[111]

Finally, we briefly outline another exciting perspective constituted by a combination of quantum and neuromorphic technologies.[112] Currently, quantum AI[113] attracts significant attention by increasingly competing with more traditional quantum computing. One of the most promising quantum AI paradigms is quantum reservoir computing,[114] which offers not only much larger state space than classical reservoir computing but also essentially non-classical quantum feedback on the reservoir via measurements. A quantum reservoir built from quantum memristors[115,116] could significantly gain quantum AI efficiency as it can readily be integrated with existing quantum and classical AI devices and also lead to an "exponential growth"[117] in the performance of "reservoirs" with the possibility of relaxing requirements on decoherence compared to traditional quantum computing.

The above trends and directions in the development of the primitives for neuromorphic computing are obviously only a slice of exciting future AI hardware technology. Even though we recognize that our choice is subjective, we hope that the outlined systems should provide a flavor of future computational hardware, which should be based on reconfigurable life-mimicking devices

utilizing different physical principles in combination with novel mathematical cognitive paradigms.[118–121]

IV. MATURE TECHNOLOGIES (COMPUTING APPROACHES)

A. SRAM

Nitin Chawla and Giuseppe Desoli

1. Status

SRAM-based computing in memory (CIM) or in-memory computing is seen as a mature and widely available technology for accelerating matrix and vector calculations in deep learning applications, yet many technology driven optimizations are still possible. To make CIM more compatible, researchers have been exploring ways to improve the design of the bitcell (Fig. 6), which is the basic unit of memory. This has led to the development of high-end SRAM chips with large capacities, such as 107, 128, and 256 Mb SRAM chips at 10, 7, 5, and 4 nm.[122–125] These large SRAM capacities help reduce the need for off-chip DRAM access. However, in more cost-sensitive applications, such as embedded systems and consumer products, modifying the bitcell design can be too costly and may limit the ability to easily transfer the technology to different manufacturing nodes.

A key difference exists between analog and digital SRAM CIM. Analog CIM has been heavily studied using capacitive or resistive sharing techniques to maximize row parallelism,[126] but this comes at the cost of inaccuracies and loss of resolution due to variations in devices across process, voltage, and temperature (PVT) and the limitations of signal-to-noise ratio (SNR) and dynamic range in analog-to-digital converter (ADC)/readout circuits. The impacts of device variations for different kinds of devices are listed below.

- Resistive devices, such as PCM or RRAM, experience a variation in the resistive values across the nominal behavior, which can vary based on the process, and for a case of ±10%–20% change in resistance value, there will be a corresponding change of current values, which are then input to the readout circuits, and hence, this will impact the quantization step of readout circuits, hence impacting the SNR, which will then need a higher dynamic range to compensate for the same. The temperature behavior for resistors also needs to be taken care in the noise margin.
- MOS devices: These devices can vary in their performance (threshold voltage) due to the following:
  1. Global lot positioning, such as slow, typical, and fast, can vary around ±20%, which can be less or more based on technology and voltage of operation. This is a deterministic shift.
  2. Local variation: within the same lot, there are device to device variations, which are random in nature and need statistical analysis based on capacity in use to analyze the impact of variations. These impact the SNR and quantization as in the case of resistive devices and will need a higher dynamic range to compensate for the loss in accuracy.

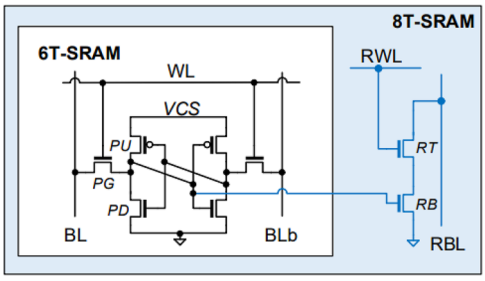

FIG. 6. Standard SRAM bitcells are usually designed with six or eight transistors.

Analog SRAM CIM solutions often use large logic bitcells and an aggressive reduction in ADC/readout bit width, resulting in low memory density and computing inaccuracies, making it difficult to use in situations where functional safety, low-cost testing, and system scalability are required. On the other hand, digital CIM offers a fast path for the next generation of neural processing systems due to its deterministic behavior and compatibility with technology scaling rules.

Researchers have improved the SRAM-based CIM's performance by modifying the SRAM bitcell structure and developing auxiliary peripheral circuits. They proposed read–write isolation cells to prevent storage damage and transposable cells to overcome storage arrangement limitations. Peripheral circuits, such as digital-to-analog converters (DACs), redundant reference columns, and multiplexed ADCs, were proposed to convert between analog and digital signals. The memory cell takes up most of the SRAM area in the core module of a standard SRAM cut. However, the complexity of the additional operations performed in the memory unit poses additional problems to utilize the memory cells to their full potential. Researchers have explored various trade-offs to implement the necessary computational functionality while preserving density and power and, last but not least, minimizing the additional cost associated with bitcell modifications required for requalification when deployed in standard design flows. Most system-on-chips (SoCs) use standard 6T structures due to their high robustness and access speed and to minimize area overhead. The 6T storage cell is made up of two P-channel Metal-Oxide-Semiconductor (PMOSs) and four N-channel Metal-Oxide-Semiconductor (NMOSs) to store data stably. To perform CIM using the conventional 6T SRAM cell, operands are represented by the word line (WL) voltage and storage node data, and processing results are reflected by the voltage difference between bit line (BL) and bit line bar (BLB).

Figure 6 shows the conventional 6T and 8T bitcells that form the basic building block of the SRAM design. The 8T bitcell is made out of a conventional 6T bitcell and a read port that allows read and write in parallel. These bitcells were never designed for parallel access across rows, and this poses one of the main challenges for enabling analog SRAM CIM.

Dual-split 6T cells with double separation have been proposed,[127,128] allowing for more sophisticated functions due to

the separated WL and GND, which can use different voltages to represent various types of information. Dong *et al.*[129] proposed a 4 + 2T SRAM cell to decouple data from the read path. The read is akin to that of the standard 6T SRAM, writes instead, and use the N-well as the Word-Line Write (WWL) and two PMOS sources as the Write-Bit Line (WBL) and Write Bit Line Bar (WBLB). In computational mode, different voltages on the WL and storage node encode the operands.

In general, CIM adopting the 6T bitcell structure is unable to efficiently perform computing operations and may not fully meet the requirements of future CIM architectures. Hence, many studies on CIM have modified the 6T structure because using the 6T standard cell directly poses a reliability challenge as the contents of the bitcells get effectively shorted if accessed in parallel on the same bit line. This means that special handling on the word line voltage is required, which adds lot of complexity and limits the dynamic range. Furthermore, the variability and linearity of devices become very difficult to control when limiting the device operation to reduced voltage levels due to these reliability constraints, impacting the overall energy efficiency of the solution.[130,131]

For practical applications, and specifically for AI ones, it is important to evaluate the end to end algorithmic accuracy vs the key metrics. To this end, recent research[132–136] has suggested various analytical models to examine the balance between the costs (accuracy) and benefits (primarily, energy efficiency and performance) of digital vs analog SRAM CIM. This is based on the idea that many machine and deep learning algorithms can tolerate some degree of computational errors and that there are methods such as retraining and fine-tuning as well as hardware-aware training to address these errors.

The implementation of neural processing units incorporating CIM components for large-scale deep neural networks (DNNs) presents significant difficulties, CIM macros can incur substantial column current magnitudes, which can result in power delivery difficulties and sensing malfunctions. Furthermore, the utilization of analog domain operations necessitates the incorporation of ADCs and DACs, which consume a significant amount of area and energy resources. Further to this, the pitch matching of ADCs with SRAM bitcells also poses a big challenge for arrays and ADC interfaces. It is clear that the realization of the full potential of SRAM-CIM necessitates development of innovative and sophisticated techniques.

2. Challenges and potential solutions

In deep learning, convolutional kernels and other types of kernels rely heavily on matrix/vector and matrix/matrix multiplication (MVM). These operations are computationally expensive and involve dot product operations between activation and kernel values. In-memory multiplication in CIM macro devices can be classified into three primary categories: current-based, charge-sharing-based for analog computation, and one for all-digital. All-digital CIM exhibits the same level of precision as purely digital Application-Specific Integrated Circuit (ASIC) implementations. Various implementation topologies ranging from bit-serial to all parallel arithmetic implementations have been proposed for digital CIM solutions. Digital CIM, as in a previous study,[137] represents a modified logic bitcell to support element-wise multiplication followed by a digital accumulation tree sandwiched within

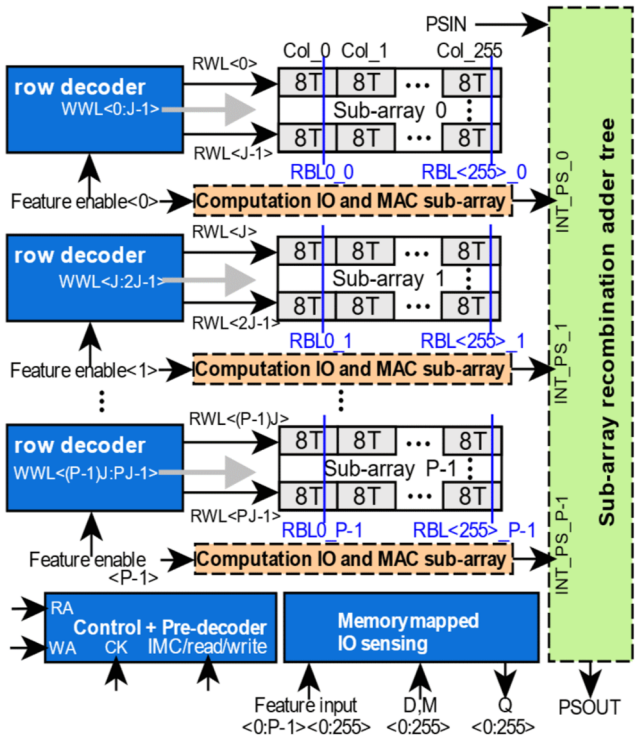

FIG. 7. Digital CIM memory macro with 8T bitcells and embedded digital logic.

the SRAM array. The solution improves on energy efficiency by reducing data movement alongside the efficiency benefits of a custom-built Multiply Acummulate (MAC) pipeline with improved levels of parallelism over traditional digital Neural Processing Units (NPUs), for example, as in Fig. 7.[138] The digital CIM implementations also have a wide voltage and frequency dynamic range allowing runtime reconfigurability between the competing Tera Operations Per Second Per Watt (TOPS/W) and TOPS/mm^2 performance criteria. The operating range and mission profiles of these architectures can also be extended by leveraging read-and-write assist schemes as is commonly done for ultra-low voltage SRAM design. The digital CIM solution's energy efficiency depends on the operand precision, and due to the deterministic precision and bit true computation nature, it begins to decline as we increase the operand precision.

Current-based CIM, as represented in one of the early research works,[139] implements a WL DAC driving a multi-level feature input with multiple rows active in parallel. The results of the element-wise multiplications of all the parallel rows are accumulated as current on the bit lines of the CIM macro that terminates in a current-based readout/ADC. The current accumulation on the bit line essentially implements a reduction operation limited by the SNR of the readout circuit. Current-based CIM, as presented in this work, suffers from significant degradation in accuracy due to bitcell variabilities and nonlinearities of the WL DAC, while the throughput is limited by the readout circuits. Kang *et al.*[140] implemented a variation with CIM using a 6T-derived bitcell with a Pulse Width Modulation (PWM) WL modulation and focused on storing and computing multi-bit weights per column. The modulation scheme uses binary weighted

pulse duration based on the index of the bitcell in the column effectively encoding the multi-bit weight in the column to impact the value of the global bit line. The multiplication is effectively done in the periphery of this CIM using a switched capacitor circuit. Bitcell variations and nonlinearities as in the previous case significantly limit the accuracy of this implementation, thus restricting the industrialization potential of these current-based CIM solutions. The work in Ref. 141 tries to address the limitations of the above analog CIM techniques and implements charge-based CIM by using a modified SRAM logic bitcell that performs element-wise binary multiplication (XNOR) and transfers the results to a small capacitor. Multiple rows operating in parallel is key to the energy efficiency of these CIM topologies. In this work, the element-wise multiplication result is transferred as a charge to the global bit line followed by a voltage-based readout. The inherent implementation benefits from the fact that capacitors suffer less from process variability and present fairly linear transfer characteristics. This architecture, however, like other analog CIM is impacted by dynamic range compression due to the limited SNR regime of the readout at the end of the column. Jia *et al.*[142] extended this approach to support multi-bit implementations using a bit-sliced architecture. The multi-bit weights are mapped to different columns, while the feature data are essentially transferred as 1-bit serial data on parallel word lines and each column performs a binary multiplication followed by accumulation on the respective bit lines. The near-memory all-digital recombination unit in this approach performs the shift and scale operations based on the column index to recreate the results of the multi-bit MAC operation. The approach is flexible to support asymmetric features and weight precision and can be made reconfigurable to support different features and weight precisions on the fly. This, however, still suffers from the same SNR constraints as each column operation is compressing the dynamic range and is limited by the peak dynamic range of the readout ADC. The ADC in most of these schemes is mostly shared across multiple columns, thus making it a critical design component in determining the throughput of such CIM architectures. The specific bit-sliced approach has impressive TOPS/W numbers for the lower weights and activation precision regime but starts to taper off due to the quadratic increase in the computation energy with increasing weight and activation precision The work in Ref. 143 instantiates multiple of these CIM macros to demonstrate a system-level approach connecting these CIM macros with a flexible interconnect and adding digital SIMD and scalar arithmetic units to support real-world neural network execution. This specific work due to the limited readout speed of the CIM macros and the overhead of the other digital units suffers from a moderate TOPS/mm^2 number for the full solution but presents impressive TOPS/W numbers, especially at the lower precision regime. The work in Ref. 144 represents another effort with a system-level solution of a hybrid NPU comprising analog CIM units and traditional digital accelerator blocks. The work leverages a low-precision (2-bit) analog CIM macro coupled with a traditional 8-bit digital MAC accelerator. The two orders of magnitude difference in energy efficiency between the 8-bit digital MAC engine and the 2-bit analog CIM macro can be leveraged by mapping different layers to the appropriate computation engine but needs careful articulation of mapping algorithms with the precision constraints of the analog CIM while keeping the overhead of the write refresh and other digital vector/scalar operators low. This, to some extent, is a trade-off between a very specialized use case and a general-purpose NN accelerator.

3. Conclusions

Analog CIM solutions based on charge-based CIM display a lower degree of variability when compared to current-based CIM, due to variability in the technologies employed for capacitors and threshold voltage effects. In addition, charge-based CIM solutions are able to activate a greater number of word lines per cycle and thus achieve higher amounts of row parallelism. However, both current-based and charge-based CIM are limited in terms of accuracy and the equivalent bit precision of the dynamic range of the accumulation. Selecting an appropriate ADC bit-precision and MVM parallelism is a challenging task that requires balancing accuracy and power consumption. Measurements and empirical evidence suggest that an increase in the accumulation values is correlated with a higher degree of variability. However, it is important to note that such high values are relatively rare in practical neural network models, as shown by the statistical distribution of activation data and resultant accumulation outcomes. This characteristic along with noise-aware training can be leveraged to optimize the precision and throughput demands of the analog-to-digital converter, thereby improving the figure of merit (FOM) of these analog CIM techniques. The research on noise-aware training in the state of the art is limited to academic works on relatively small neural networks and datasets. This for an industrial deployment still needs to mature and demonstrate scalability to larger models and datasets.

All-digital CIM provides a deterministic and scalable path to intercept the implementation of NPUs by bringing an order or more of gain vs traditional all-digital NPUs. Digital CIM solutions provide excellent scaling for area and energy efficiency as we move toward more advanced CMOS nodes with a wide operating voltage and frequency range tunability while still maintaining a general purpose and application-agnostic view of embedded neural network acceleration at the edge.

On the other hand, for applications that can handle approximate computing, analog SRAM CIM-based solutions provide a much-increased level of computation parallelism and energy efficiency while still operating in an SNR-limited regime. The impacts of dynamic range compression and readout throughput are key algorithmic and design trade-offs while designing an analog CIM solution, which tries to operate in a much more restricted voltage and frequency regime as opposed to a digital CIM solution. That the application choices are more vertically defined as opposed to general purpose is also a deciding factor in choosing an analog SRAM CIM-based solution as opposed to digital CIM solutions. In conclusion, due to the rapid industrialization potential of SRAM-based CIM solutions and the opportunity of exploiting the duality of these CIM instances to serve as SRAM capacity to support the system in other operating modes, there are enough reasons to remain invested in SRAM CIM. The scope to improve both digital and analog SRAM CIM remains very high, both at the design and at the technology level, to exploit the best gains out of these two solutions, which in the future can also be combined to form a hybrid solution serving multiple modalities of neural network execution at the edge.

B. Flash memories

Gerardo Malavena and Christian Monzio Compagnoni

1. Status

Thanks to a relentless expansion in all the application fields of electronics since their conception in the 1980s, flash memories became ubiquitous non-volatile storage media in everyday life and a source of market revenues exceeding $60B in 2021. The origin of this success can be traced back to their capability to solve the trade-off against cost, performance, and reliability in data storage much better than any other technology. Multiple solutions to that trade-off, in addition, were devised through different design strategies that, in the end, allowed flash memories to target a great variety of applications in the best possible way. Among these different design strategies, the two leading to the so-called NOR flash memories[145] and NAND flash memories[146] became by far the most important.

As in all flash memory designs, NOR and NAND flash memories store information in memory transistors arranged in an array whose operation relies on an initialization, or erase, step performed *in a flash* on a large number of devices simultaneously. In particular, the erase step moves the threshold-voltage (V_T) of all the memory transistors in a block/sector of the array to a low value. From that initial condition, data are stored through program steps performed in parallel on a much smaller subset of memory transistors, raising their V_T to one or more predefined levels. This working scheme of the array allows us to minimize the number of service elements needed for information storage and, in the end, is on the basis of the high integration density, high performance, and high reliability of flash memories. Starting from it, the structure of the memory transistors, the architectural connections among them to form the memory array, the array segmentation in the memory chip, the physical processes exploited for the erase and program steps, and many other aspects are markedly different in NOR and NAND flash memories.

NOR flash memories follow a design strategy targeting the minimization of the random access time to the stored data, reaching latencies as short as a few tens of nanoseconds. A strong array segmentation is then adopted to reduce the delay time of the word-lines (WLs) and bit-lines (BLs) driving the memory transistors. As depicted in Fig. 8(a), moreover, the memory transistors are independently connected to the WLs, BLs, and source lines (SLs) of the array to simplify and speed up the sequence of steps needed to randomly access the stored data and to allow device operation at relatively high currents (currents in the microampere scale are typical to sense the data stored in the memory transistors). Fast random access is also achieved through a very robust raw array reliability, with no or limited adoption of error correction codes (ECCs). This design strategy, on the other hand, does not make NOR flash memories the most convenient solution from the standpoint of the area and, hence, the cost of the memory chip and limit the chip storage capacity to low or medium sizes (up to a few Gbits).

NAND flash memories rely on a design strategy pointing to the minimization of the data storage cost. Therefore, limited array segmentation is adopted and the memory transistors are in series

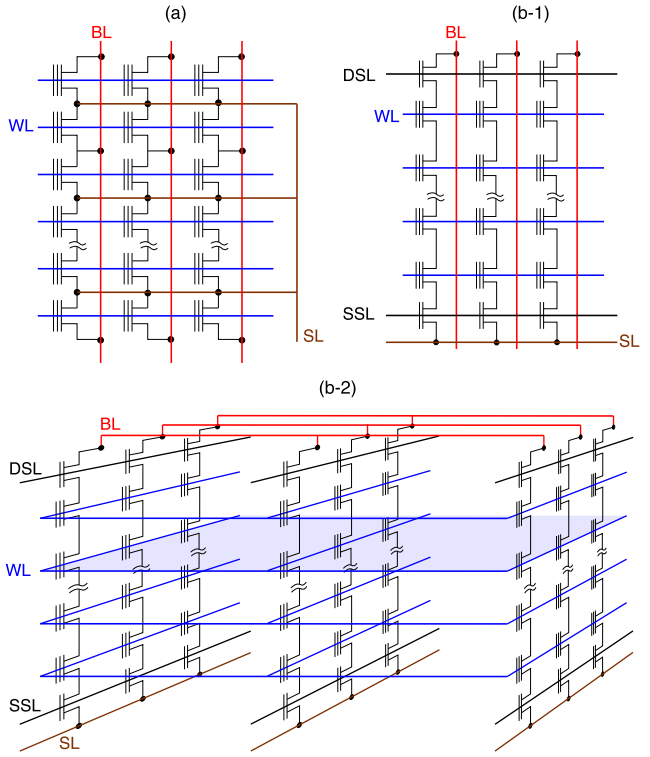

FIG. 8. Schematic for the connection of the memory transistors in (a) a NOR flash memory array (a common ground architecture of stacked-gate memory transistors has been assumed); (b-1) a planar and (b-2) a vertical (3D) NAND flash memory array.

connection along strings to reduce the area occupancy of the memory chip. Figures 8(b-1) and 8(b-2) schematically show the arrangement of the memory transistors in a planar and in a vertical (or, 3D) NAND flash array, respectively. At present, 3D arrays represent the mainstream solution for NAND flash memories, capable of pushing their bit storage density up to 15 Gbits/mm^2,[147] a level unreachable by any other storage technology. Such an achievement was made possible also by the use of multi-bit storage per memory transistor and resulted in memory chips with capacity as high as 1 Tbit.[147] The NAND flash memory design strategy, on the other hand, makes the random access time to the stored data relatively long (typically, a few tens of microseconds). That is the outcome of time delays of the long WLs and BLs in the microsecond timescale, low sensing currents (tens of nanoampere) during data retrieval due to series resistance limitations in the strings, and the need of multi-bit detection per memory transistor. In addition, array reliability relies on powerful ECCs.

Given the successful achievements of flash memories as non-volatile storage media for digital data, exploiting them in the emerging neuromorphic-computing landscape appears as a natural expansion of their application fields and is attracting widespread interests. In this landscape, flash memories may work not only as storage elements for the parameters of artificial neural networks (ANNs) but also as active computing elements to overcome the von Neumann

bottleneck of conventional computing platforms. The latter may represent, of course, the most innovative and disruptive application of flash memories in the years to come. At the same time, the use of flash memories as active computing elements may boost the performance, enhance the power-efficiency, and reduce the cost of ANNs, making their bright future even brighter. In this context, relevant research has been focusing on employing flash memory arrays as artificial synaptic arrays in hardware ANNs and as hardware accelerators for the vector-by-matrix multiplication (VMM), representing the most common operation in ANNs. Quite promising results have already been reported in the field, through either NOR[148,149] or NAND[150–155] flash memories. In these proofs of concept, different encoding schemes for the inputs (e.g., voltage amplitude or pulse width modulation, with signals on the BLs or WLs of the memory array) and different working regimes of the memory transistors have been successfully explored. Interested readers may go through the references provided in this section for a detailed description of the most relevant schemes proposed so far to operate a flash array as a computing element.

2. Challenges and potential solutions

Despite the encouraging proofs of concept already reported, the path leading to flash memory-based ANNs still appears long and full of challenges. The latter can be classified into the following categories:

a. Challenges arising from changes in the design strategy of the array. As previously mentioned, the success of flash memories as non-volatile storage media for digital data arises from precise design strategies. Modifying those strategies to meet the requests of ANNs may deeply impact the figures of merit of the technology and should be carefully done. For instance, ANN topologies requiring to decrease the segmentation of NOR flash arrays may worsen their performance in terms of working speed. Increasing the segmentation of NAND flash arrays to meet possible ANN topology constraints or to enhance their working speed may significantly worsen their cost per memory transistor.

The cost per memory transistor of flash memories, in addition, is strictly related to the array capacity. Modifying the latter or not exploiting it all through the ANN topology may reduce the cost-effectiveness of the technology. In this regard, the very different capacities of NOR and NAND flash arrays make the former suitable for small/medium size ANNs (less than 1 giga parameters) and the latter suitable for large size ANNs (more than 1 giga parameters). The organization of the memory transistors into strings in NAND flash arrays represents an additional degree of complexity for the exploitation of their full capacity in ANNs. In fact, the number of memory transistors per string is the outcome of technology limitations and cost minimization and, therefore, cannot be freely modified. Exploiting all the memory transistors per string, then, necessarily sets some constraints on the ANN topology (the number of hidden layers, the number of neurons, etc.), which, of course, should be compatible with the required ANN performance.

Another important aspect to consider is that the accurate calibration of the V_T of the memory transistors needed by high-performance and reliable ANNs may not be compatible with the block/sector erase scheme representing a cornerstone of all the design strategies of flash memories. Solutions to carry out the erase step on single memory transistors are then to be devised. These solutions may require a change of the array design as in Refs. 148 and 149 or new physical processes and biasing schemes of the array lines to accomplish the erase step as in Refs. 156–158. All of these approaches, however, necessarily impact relevant aspects of the technology, affecting its cost, performance, or reliability, and should be carefully evaluated.

The change of the typical working current of the memory transistors when exploiting flash memories for ANN applications is another critical point to address. In fact, reducing the working current of the memory transistors may make it more affected by noise and time instabilities. Increasing it too much, on the other hand, may raise issues related to the parasitic resistances of the BLs and SLs, and, in the case of NAND flash arrays, of the unselected cells in the strings.

b. Challenges arising from array reliability. Flash memories are highly reliable non-volatile storage media for digital data. That, however, does not assure that they can satisfactorily meet the reliability requirements needed to operate as computing elements for ANN applications. Especially in the case of NAND flash memories, in fact, array reliability in digital applications is achieved through massive use of ECCs and a variety of smart system-level stratagems to take under control issues, such as electrostatic interference between neighboring memory transistors, lateral migration of the stored charge along the charge-trap storage layer of the strings, and degradation of memory transistors after program/erase cycles. All of that can hardly be exploited to assure the reliable operation of flash arrays as computing elements. In addition, the requirements on the accuracy of the placement and the stability over time of the V_T of the memory transistors when using flash arrays as computing elements may be more severe than in the case of digital data storage. The possibility to satisfy those requirements in the presence of the well-known constraints to the reliability of all flash memory designs[146,159,160] is yet to be fully demonstrated. In this context, periodic recalibration of the V_T of the memory transistors and on-chip learning[155] may mitigate the array reliability issues.

c. Challenges arising from the peripheral circuitry of the array. As in the case of flash memory chips for non-volatile storage of digital data, the peripheral circuitry of flash memory arrays used as computing elements for ANNs should not introduce severe burdens on the chip area, cost, power efficiency, and reliability. In the latter case, this aspect is particularly critical due to the need to integrate on the chip not only the circuitry to address the memory transistors in the array and to carry out operations on them, but also, for instance, the circuitry to switch between the digital and the analog domain in VMM accelerators or to implement artificial neurons in hardware ANNs. Along with effective design solutions at the circuit level,[150] process solutions, such as CMOS-under-array integration[147] or heterogeneous integration schemes,[152] should be exploited for successful technology development.

3. Conclusion

Flash memories may play a key role in the neuromorphic-computing landscape. Expanding their fields of application, they can be the elective storage media for ANN parameters. However, they can also be active computing media for high-performance, power-efficient, and cost-effective ANNs. To achieve this intriguing goal,

relevant challenges must be faced from the standpoint of the array design, reliability, and peripheral circuitry. Winning those challenges will be a matter of engineering and scientific breakthroughs and will pave the way for years of unprecedented prosperity for both flash memories and ANNs.

V. EMERGING TECHNOLOGIES (COMPUTING APPROACHES)

Zhongrui Wang and J. Joshua Yang

A. Resistive switching and memristor

1. Status

Resistive switches (often called memristors when device nonlinear dynamics are emphasized) are electrically tunable resistors, of a simple metal–insulator–metal structure. Typically, their resistance changes as a result of redox reactions and ion migrations, driven by electric fields, chemical potentials, and temperature.[64] There are two types of resistive switches according to the mobile ion species. In many dielectrics, especially transition metal oxides and perovskites, anions such as oxygen ions (or equivalently oxygen vacancies) are relatively mobile and can form a conduction percolation path, leading to the so-called valence change switching. For example, a conical pillar-shaped nanocrystalline filament of the Ti_4O_7 Magnéli phase was visualized using a transmission electron microscope (TEM) in a $Pt/TiO_2/Pt$ resistive switch.[161] On the other hand, the conduction channels can also be created by the redox reaction and migration of cations, which involves the oxidation of an electrochemically active metal, such as Ag and Cu, followed by the drift of mobile cations in the solid electrolyte and the nucleation of cations to establish a conducting channel upon reduction. The dynamic switching process of a planar Au/SiO_x:Ag/Au diffusive resistive memory cell was captured by *in situ* TEM.[89]

Resistive switches provide a hardware solution to address both the von Neumann bottleneck and the slowdown of Moore's law faced by conventional digital computers. When these resistive switches are grouped into a crossbar array, they can naturally perform vector–matrix multiplication, one of the most expensive and frequent operations in machine learning. The matrix is stored as the conductance of the resistive memory array, where Ohm's law and Kirchhoff's current law physically govern the multiplication and summation, respectively.[64] As a result, the data are both stored and processed in the same location. This in-memory computing concept can largely obviate the energy and time overheads incurred by expensive off-chip memory access on conventional digital hardware. In addition, the resistive memory cells are of simple capacitor-like structures, equipping them with excellent scalability and 3D stackability. So far, resistive in-memory computing has been used for hardware implementation of deep learning models to handle both unstructured (e.g., general graphs, images, audios, and texts) and structured data, as discussed in the following.

General graph: Graph-type data consist of a set of nodes together with a set of edges. The theoretical formulation has been made for graph learning using resistive memory on datasets such as WikiVote.[162,163] Experimentally, a resistive memory-based echo state graph neural network has been used to classify graphs in MUTAG and COLLAB datasets as well as nodes in the CORA dataset,[164] including few-shot learning of the latter.[165]

Images: Images are special graph-type data. Both supervised and unsupervised learning of ordinary images have been experimentally implemented on resistive memory. For supervised learning, offline trained resistive memory, where optimal conductance of memory cells is calculated by digital computers and transferred to resistive memory, is used to classify simple patterns,[166,167] MNIST handwritten digits,[168–171] CIFAR-10/100 datasets,[172–174] ImageNet,[175] and Omniglot one-shot learning dataset.[176] In addition to offline training, online training adjusts the conductance of resistive memory in the course of learning, which is more resilient to hardware nonidealities in classifying simple patterns,[37,166] Yale Face and MNIST datasets,[177,178] CIFAR-10 dataset,[179] and meta-learning of Omniglot dataset.[180] Besides supervised learning, unsupervised offline learning with resistive memory is used for sparse coding of images[181] and MNIST image restoration.[182]

Audios and texts: Learning sequence data, such as audios and texts, have been implemented on resistive memory. Supervised online learning using recurrent nodes has been done on the Johann Sebastian Bach chorales dataset.[183] In addition, delayed-feedback systems based on dynamic switching of resistive memory are used for temporal sequence learning, such as spoken number recognition and chaotic series prediction.[66,184,185] For offline learning, resistive memory is used for modeling the Penn Treebank dataset;[186] Wortschatz Corpora language dataset and Reuters-21578 news dataset;[187] and Bonn epilepsy electroencephalogram dataset and NIST TI-46 spoken digit dataset.[188,189]

Structured data: Despite unstructured data, structured data, such as those of a tabular format, have been tackled by resistive memory, including supervised classification of the Boston housing dataset on an extreme learning machine;[190] K-means clustering of IRIS dataset and principal component analysis of the breast cancer Wisconsin (diagnostic) dataset;[191,192] and correlation detection of quality controlled local climatological database.[193]

2. Challenges

Major challenges can be categorized at different levels.

Device level: The ionic nature of resistive switching, although benefits data retention, imposes challenges on programming precision, energy, and speed. The programming precision limits the representation capability of the resistive switch, or equivalently how many bits a device can encode. In addition, the programming energy and speed impact online learning performance. In addition, the degradation of the representation capability is further intensified by the read noise, manifestation by the current fluctuation under a constant voltage bias.

Circuit level: Analog resistive memory arrays are mostly interfaced with up- and downstream digital modules in a computing pipeline. As such, there is inevitable signal acquisition and conversion cost, which leads to the question of how to achieve trade-off between signal acquisition rate, precision, and power consumption. In addition, the parasitic resistance and capacitance, like the non-

zero wire resistance, incur the so-called IR drop in the resistive memory crossbar array.

Algorithm level: So far, many applications of resistive memory suffer from significant performance loss in the presence of resistive memory nonlinearities (e.g., noises), thus defeating their efficiency advantage over alternative digital hardware.

3. Potential solutions

Device level: Various approaches are used to address the programming stochasticity, such as the local confinement of conducting filament.[194] A denoising protocol using sub-threshold voltages has recently been developed to suppress the fluctuation of the device state and achieve up to 2048 conductance levels.[195] In addition, homogeneous switching may suppress stochasticity at the cost of larger program energy and time overheads.[64] In terms of programming energy, small redox barriers and large ion mobilities may reduce switching energy and accelerate switching speed, at the expense of retention and thermal stability though.

Circuit level: Typically, resistive in-memory computing relies on Ohm's law and Kirchoff's current law, resulting in current summation. However, there is a recent surge of interest in replacing current summation by voltage summation, which lowers down the static power consumption by eliminating current summation incurred Joule heating. In addition, fully analog neural networks have been proposed to get rid of the frequent analog-to-digital and digital-to-analog conversions.[196] To combat with the parasitic wire resistance, a simple solution is to increase device resistance in both ON and OFF states, such as that demonstrated in a 256 × 256 in-memory computing macros.[195]

Algorithm level: A recent trend is hardware–software co-design to leverage resistive memory nonlinearities and turn them into advantages. For example, the programming stochasticity can be exploited by neural networks of random features (e.g., echo state networks[164,165] and extreme learning machines[190]) and Bayesian inference using Markov Chain Monte Carlo (MCMC), such as Metropolis–Hastings algorithm.[197] In addition, such programming noise is a natural regularization to suppress overfitting in online learning.[198] Moreover, hyperdimensional computing[187] and mixed-precision design, such as high-precision iterative refinement algorithm paired with low-precision conjugate gradient,[199] can withstand resistive memory programming noise. The reading noise can also be exploited for solving combinatorial optimization problems using simulated annealing, serving as a natural noise source to prevent the system from falling into the local minimum.[200,201]

4. Conclusion

The advent of resistive switch-based in-memory computing in the past decade has demonstrated a wide spectrum of applications in machine learning and neuromorphic computing, reflected by its handling of different types of data.

However, there is still plenty of room, at device, circuit, and algorithm levels, to improve, which will help fully unleash the power of in-memory computing with resistive switches and potentially yield a transformative impact on future computing.

B. Phase change materials

Abu Sebastian and Ghazi Sarwat Syed

1. Introduction

Phase-change memory (PCM) is arguably the most advanced memristive technology. Similar to conventional metal-oxide based memristive devices, information is stored in terms of changes in atomic configurations in a nanometric volume of material and the resulting change in resistance of the device.[202] However, unlike the vast majority of memristive devices, PCM exhibits volumetric switching as opposed to filamentary switching. The volumetric switching is facilitated by certain material compositions along the $GeTe–Sb_2Te_3$ pseudo-binary tie line, such as $Ge_2Sb_2Te_5$, that can be switched reversibly between amorphous and crystalline phases of different electrical resistivities.[203] Both transitions are Joule-heating assisted. The crystalline to amorphous phase transition relies on a melt-quench process, whereas the reverse transition relies mostly on crystal growth (Fig. 9).

There are essentially two key properties that make PCM devices particularly well suited for neuromorphic computing[204] (see Fig. 10). Interestingly, this was pointed out by Stanford Ovshinsky, a pioneer of PCM technology, way back in 2003 when PCM was being considered just for memory applications.[205] The first property is that PCM devices can store a range of conductance values by modulating the size of the amorphous region typically achieved by partial RESET pulses that melt and quench a certain volume of the PCM material. This analog storage capability, combined with a crossbar topology, allows for matrix–vector multiplication (MVM) operations to be carried out in O(1) time complexity by leveraging Kirchhoff's circuit laws. This makes it possible to realize an artificial neural network on crossbar arrays of PCM devices, with each synaptic layer of the DNN mapped to one or more of the crossbar arrays.[67,206] The second property referred to as accumulative property results from the progressive crystallization of the PCM material upon application of an increasing number of partial SET pulses. It is used for implementing DNN training;[207] temporal correlation detection;[193] continual learning;[208] local learning rules, such as spike-timing-dependent plasticity;[209,210] and neuronal dynamics.[211]

PCM is at a very high maturity level of development and has been commercialized as both stand-alone memory[212] and embedded memory.[213] This fact, together with the ease of embedding PCM on logic platforms (embedded PCM),[206] makes this technology of unique interest for neuromorphic computing.

2. Challenges

PCM devices offer write operations in the tens of nanosecond timescale, which is sufficient for most neuromorphic applications, in particular those targetting deep learning inference. The cycling endurance could also exceed a billion cycles (dependent on the device geometry), which is several orders of magnitudes higher than commercial flash memory.[214] This is sufficient for deep learning inference applications. The cycling endurance for partial SET pulses is much higher than that for full SET–RESET cycling and hence is widely considered sufficient for other neuromorphic applications as

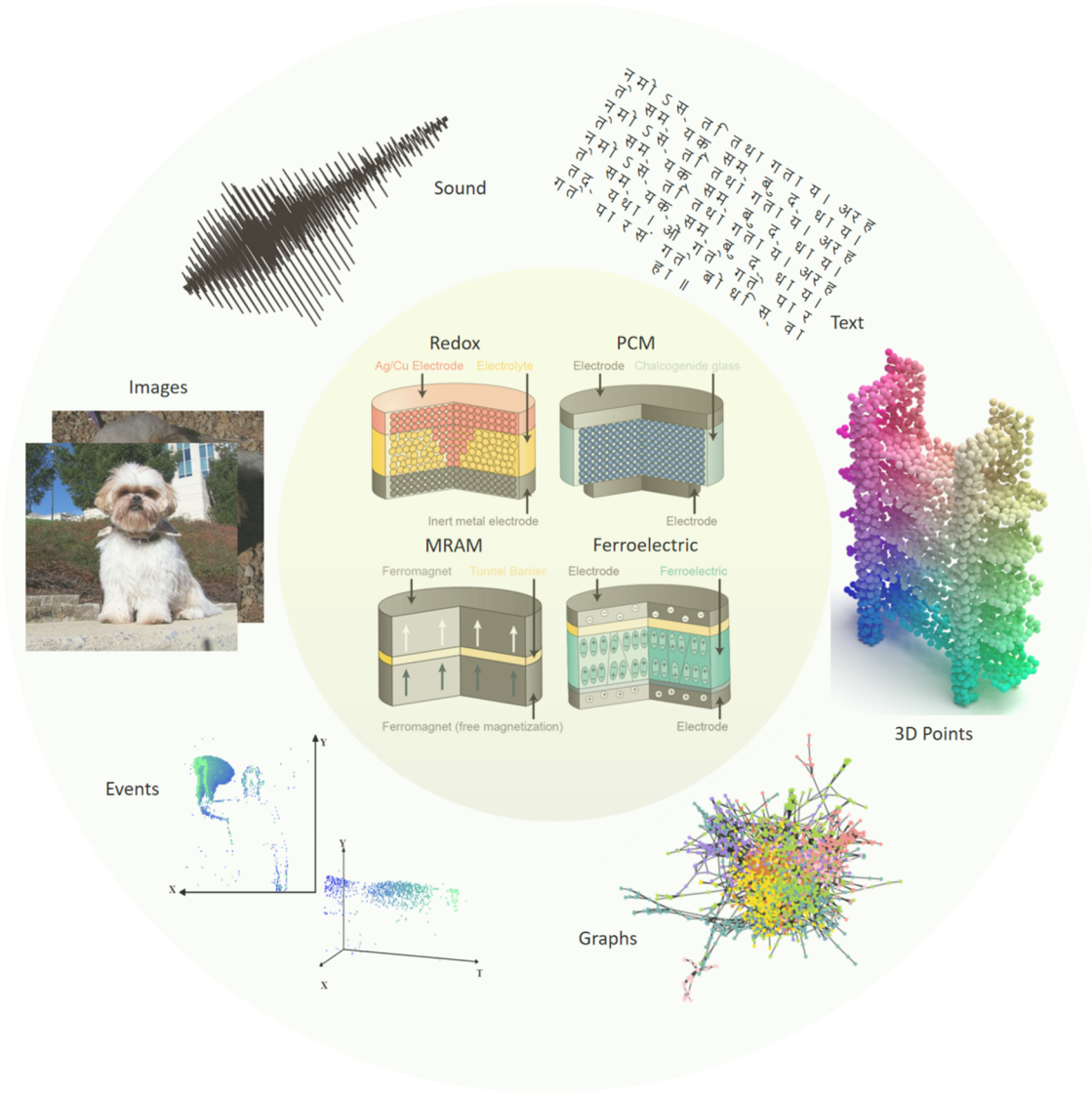

FIG. 9. Summary of emerging memories, such as memristors, and their capabilities in processing various data types, such as images, audios, texts, 3D points, graphs, and events. Image samples are taken from the ImageNet dataset. The audio waveform visualizes a sample from the TIDIGITS dataset. 3D points visualize a sample from the ModelNet10 dataset. The graph sample is from the CORA dataset.

well. The read endurance is almost infinite for PCM when sufficiently low read bias is applied. Another key attribute is retention, which is typically tuned through material choice.[215] However, the use of analog conductance states in neuromorphic computing makes the retention time of intermediate phase configurations even more important, which could be substantially lower than that of fully RESET states.

One of the primary challenges for PCM is integration density. For example, for DNN inference, it is desirable to have at least $10–100 \times 10^6$ on-chip weight capacity. The crossbar array for neuromorphic computing comprises metal lines intersected by synaptic elements, which are composed of one or more PCM devices and selector devices. Access devices, such as bipolar junction transistors or metal oxide–semiconductor field effect transistors, are preferred for accurate programming, while two-terminal polysilicon diodes offer scalability. To achieve high memory density, stacking multiple crossbar layers vertically is beneficial. Back-end of-line (BEOL) selectors, such as ovonic threshold switches, show promise but face challenges in achieving precise current control. Edge effects and thermal crosstalk between neighboring cells become significant at smaller feature sizes.[247–249]

Compute precision is a crucial aspect especially for DNN inference applications. The key challenges are 1/f read noise and conductance drift[216] (see Fig. 11). Drift is attributed to the

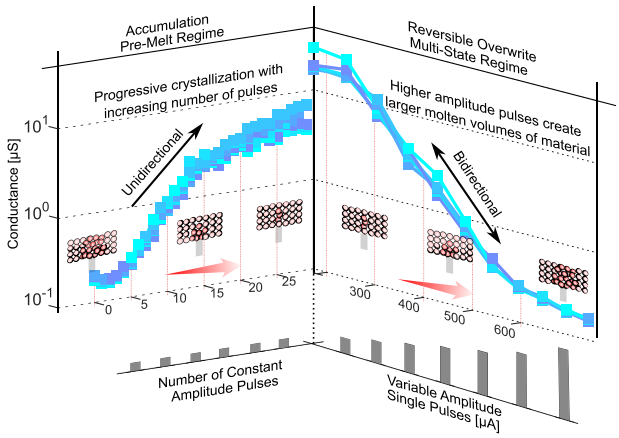

FIG. 10. Operational regimes of a phase change device when used for neuromorphic computing. On the right plot, the direct overwrite regime utilizing melt-quench dynamics is illustrated. The programming curves display the achievable conductance values in response to partial RESET pulses of varying amplitudes. As the RESET pulse is increased in amplitude, a larger amorphous volume is created mostly independent of the phase configuration prior to the application of the pulse. On the left plot, the characteristic accumulative property is demonstrated. It showcases the evolution of conductance values over successive applications of a SET pulse with a constant amplitude. As the amorphous region reduces in size due to crystallization dynamics, the device conductance progressively increases. Multiple experimental traces are overlaid in both plots.

structural relaxation of the melt-quenched amorphous phase and exhibits a log time dependence. Conductance variations arising from temperature variations could also impact the compute precision.[217] Another potential source of imprecision is voltage polarity dependence.[218] The intrinsic stochasticity associated with the accumulative behavior could also be a challenge for applications such as online learning.[219]

For applications that exploit the accumulative behavior, there is a significant incentive to minimize the programming current. In fact, reducing the programming currents could also help with achieving better integration density by minimizing the requirements on the access devices. The primary way to achieve lower programming current is via scaling down the volume of switching material. PCM devices have decreased in programming energies by a factor of 1000 since the first memory chip was reported. Some device structures now exhibit programming energies in the tens of femtojoules (i.e., on par with the most efficient charge-based memories) via extreme volume scaling.[220–223] However, analog capability is typically compromised, and extreme scaling also leads to fabrication challenges.

3. Potential solutions

Two main approaches have been taken to improve PCM devices: material engineering and device engineering. Material engineering involves exploration of new phase-change material compositions as well as alloying of phase-change materials with elements such as germanium, silicon, carbon etc.[215] Yet another approach is the use of superlattice heterostructures.[224,225] They utilize alternating layers of two different phase change materials that are only a few atoms thick, which create an electro-thermal confinement effect that enhances write efficiency.[226,227] This approach also improves the write endurance and reduces the resistance drift and noise.[228] However, additional research is necessary to fully comprehended the mechanisms and examine the impact of device geometries and the randomness that accompanies crystal growth and amorphization.[229–232]

Device engineering involves creating devices such as projected phase-change memory, which have a noninsulating projection segment that is placed in parallel to the phase-change material segment.[233,234] Another fascinating approach is that of relying on nanoscale confinement of simple materials such as antimony to design better PCM devices.[235] Besides improving the PCM devices themselves, one could also conceive innovative synaptic units with more than one PCM device to enhance the conductance window and to improve the compute precision.[236] There is also potential to enhance the compute precision by programming approaches such as gradient-descent programming that relies on minimizing the MVM error as opposed to minimizing the programming error per device.[237]

Phase-change materials have functional properties in the optical domain that enable neuromorphic computing on photonic integrated circuits using photonic phase-change memory devices.[238] By integrating these materials onto silicon waveguides,[239,240] analog multiplication of incoming optical signals becomes feasible. Additionally, spike aggregation and convolution operations can be conducted in a single time step using wavelength division multiplexing.[241,242] The accumulative behavior of phase-change materials also allows for more intricate operations such as correlation detection with high efficiency.[243] This opens opportunities for the development of novel phase-change materials engineered specifically for photonic applications.

There are also reports of PCM device non-idealities being exploited for computational purposes. For example, the stochasticity associated with the accumulative behavior can create biorealistic randomly spiking neurons,[211] and structural relaxation can be used to implement eligibility traces for reinforcement learning.[244] The conductance fluctuations in PCM have also been exploited in an in-memory factorizer to disentangle visual attributes.[245] Finally, the ability to induce field effect modulation in PCM devices combined with the analog storage capability can be exploited to realize mixed synaptic plasticities for solving optimization and sequential learning problems.[246]

4. Conclusion

With well-understood device physics models, established manufacturability, and proven integration capability with state-of-the-art CMOS logic platforms using BEOL processing, PCM becomes arguably the most advanced memristive technology. More recently PCM has been extensively researched for neuromorphic computing by exploiting its analog storage capability and accumulative behavior. However, commercialization of such technology requires improvements in achievable compute precision, integration density, all within the purview of BEOL compatible materials and processing. Moreover, as with commercialization of any emerging technology, a key deciding factor would be the manufacturing cost. The expectation is that the manufacturing cost barrier when PCM is used for computing applications is not as limiting as for storage-class

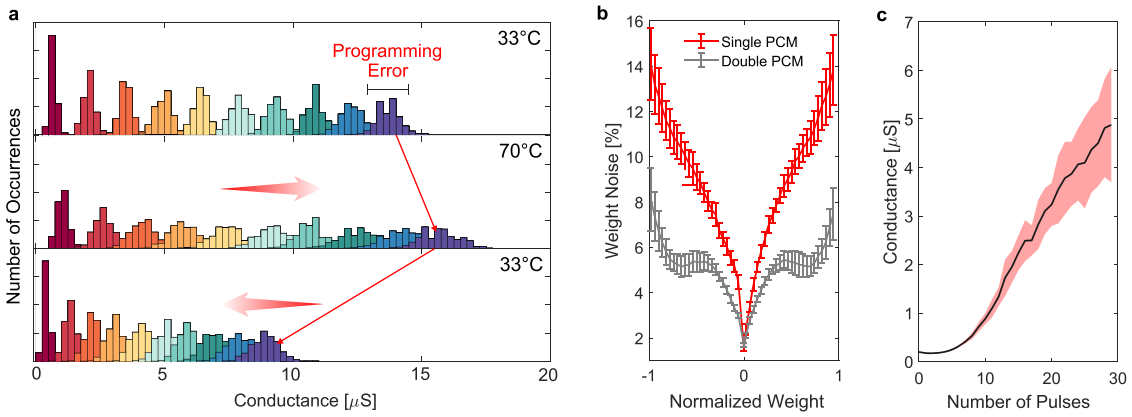

FIG. 11. PCM non-idealities. (a) Device data after programming show variability, reflected in broad distributions of analog conductance values due to programming inaccuracies, read noise, and drift variability (the top panel). Temperature increase raises state conductivity due to thermal carrier excitation and accelerates structural relaxation. (b) The conductance fluctuations manifest as synaptic weight noise, here shown as additive noise in terms of the percentage of the maximum synaptic weight. Using two PCM devices per synapse reduces this error. (c) The accumulative behavior exhibits significant stochasticity mostly attributed to variations in the crystallization kinetics.

memory applications. Most likely, neural processing units for DNN inference based on embedded PCM for analog in-memory computing would be commercialized first. Depending on the level of commercial acceptance of this technology, full-edged PCM-based accelerators could be developed to serve high-end edge applications or even cloud-based applications.

C. Ferroelectric materials

Thomas Mikolajick, Stefan Slesazeck, and Beatriz Noheda

1. Status

Ferroelectric materials are, in theory, ideally suited for information storage tasks since their switching is purely field-driven, holding the promise of extremely low write energy, and non-volatile at the same time. Moreover, unlike competing concepts, such as resistive switching or magnetic switching, ferroelectric materials offer three different readout possibilities giving a lot of flexibility in device design.[250] In detail, the following read schemes can be applied (see also the middle part of Fig. 12):

- Direct sensing of the switched charge during polarization reversal, as used in the ferroelectric RAM (FeRAM) concept, results in a cell design similar to a dynamic random-access memory (DRAM).[251]
- Coupling of the ferroelectric to the gate of a field effect transistor and readout of the resulting drain current, as used in the ferroelectric field effect transistor (FeFET). This results in a cell that is similar to classical transistor-based charge storage (floating gate or charge trapping) memory cell, which is most prominently used in flash memories.[252]
- Modulation of the tunneling barrier in a ferroelectric tunneling junction (FTJ). As a result, we can realize a two-terminal device, which is essentially a special version of a resistive switching memory cell (see Sec. V A).[253]

Each of the mentioned readout schemes has advantages and disadvantages, and therefore, the flexibility to use one of the three is a plus, especially in applications that go beyond pure memories, such as neuromorphic computing.

However, traditionally, ferroelectricity was only experienced in chemically complex materials, such as lead-zirconium titanate (PZT), strontium bismuth tantalate (SBT), or bismuth ferrite (BFO), which all are very difficult to incorporate into the processing flow for integrated electronic circuits, due to their limited stability in reducing environments. Another pervasive issue for the integration of ferroelectrics is their tendency to depolarize upon downscaling, an issue that is accentuated by their high permittivity. Organic ferroelectrics, the most prominent example being polyvinylidene difluoride (PVDF), can mitigate this problem, as their low permittivities reduce the depolarization fields, while a rather high coercive field increases the stability of the polarization state. Such materials are ideally suited for lab scale demonstrations of new device concepts, due to their simple fabrication using a solution-based process, and are highly preferred for flexible and biocompatible electronics.[254] However, their limited thermal stability has taken them out of the game for devices in integrated circuits. Therefore, although the technology in the form of FeRAM[255] is on the market for more than 25 years, it has lacked the ability to scale in a similar manner as conventional memory elements and, therefore, it is still limited to niche applications that require a high rewrite frequency together with non-volatility as in data logging applications.

2. Challenges

With the discovery of ferroelectricity in hafnia (HfO_2) and zirconia (ZrO_2), the biggest obstacle of the limited compatibility with integrated circuit fabrication could be solved.[256] HfO_2 and

ZrO_2 are stable both in reducing ambient and in contact with silicon, and their fabrication using established atomic layer deposition processes is standard in modern semiconductor process lines. However, new difficulties, especially with respect to reliability,[257] need to be solved. Challenges in this direction are aggravated by the metastable nature of the ferroelectric phase, which appears mostly at the nanoscale, making a full understanding of the polar phase quite demanding. While their high coercive field makes them very stable with respect to classical retention, the ferroelectric phase typically exists together with other non-polar phases, which prevents them from reaching the predicted polarization values (of the order of 50 $\mu C/cm^2$).[258] Moreover, the most serious problem of any non-volatile ferroelectric device, the imprint, becomes very complex to manage in hafnia/zirconia-based ferroelectrics. The imprint is a shift of the hysteresis loops due to an internal bias. While this effect leads to a classical retention of the stored state that may look perfect, after switching, retention will be degraded and fixing the so-called opposite-state retention loss needs to be carefully done by material and interface engineering. Moreover, the high coercive field in this material class becomes a problem as HfO_2 and ZrO_2 often show a pronounced wake-up and fatigue behavior and the field-cycling endurance is in many cases limited by the dielectric breakdown of the material.

While the issues mentioned so far are valid for any non-volatile device application, in neuromorphic systems, additional challenges arise, including the linearity of the switching behavior and tuning of the retention to achieve both short-term and long-term plasticity, as well as specific effects to mimic neurons, such as accumulative switching,[250,259] which need to be explored using material and device design measures. Finally, large-scale neuromorphic systems will require a high integration density that demands three-dimensional integration schemes, realized either by the punch-and-plug technology well-known from NAND flash or by integrating devices into the back-end of the line.

3. Potential solutions

Since the original report on ferroelectricity in hafnium oxide,[256] the boundary conditions for stabilizing the ferroelectric phase have been much better understood, although there are still a number of open questions. The goal is to achieve a high fraction of the ferroelectric phase without dead layers of non-ferroelectric phases at the interface to the electrodes or in the bulk of the film. This needs to be done under the boundary conditions of a realistic fabrication process, which means that sophisticated methods to control the crystal structure based on epitaxial growth are not possible. Epitaxial growth can help clarify scientific questions, but the achieved results need to be transferred to chemical vapor deposition (CVD), including most prominently atomic layer deposition (ALD), or physical vapor deposition (PVD) processes using electrodes such as TiN or TaN that can be integrated into electronic processes.

In the past years, it became obvious that oxygen vacancies are, on the one hand, required to stabilize the ferroelectric phase[260] and,

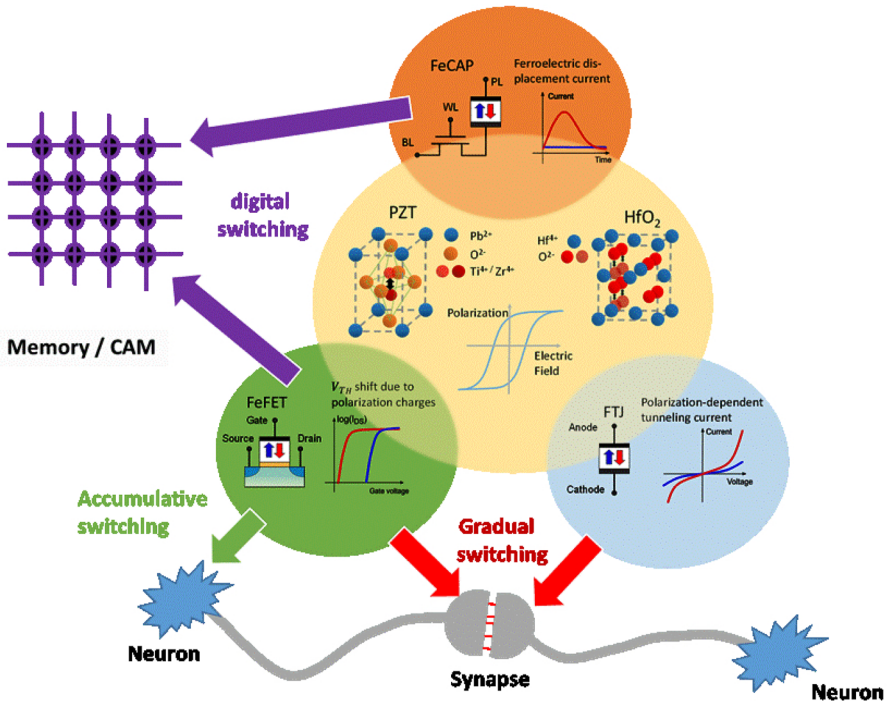

FIG. 12. Ferroelectric materials (center) enable three different basic memory cells (middle ring). These can be used in various ways in neuromorphic circuits (for examples, see the outer part of the figure). The rich switching dynamics of ferroelectrics allow us to tailor new devices mimicking neurons and synapses in a much more flexible and area efficient way as compared to their pure CMOS counterparts.

on the other hand, detrimental to both the imprint and the field cycling behavior.[261] Therefore, many proposals to integrate the ferroelectric layer with additional thin layers in the film stack have been made, and currently, a lot of work is going in that direction. Moreover, it is clear that the interface to the electrodes needs careful consideration. In this direction,[262] facilitating the transport of oxygen, not only in the ferroelectric layer but also across the electrode interfaces, by minimizing the strain effects, may be key to improving device performance.[263] When it comes to structures that are in direct contact with silicon, a recent observation of a quasi-epitaxial growth of extremely thin hafnium-zirconium oxide films on silicon could be an interesting direction.[264] For concrete neuromorphic applications, the rich switching dynamics can be very helpful (see Fig. 12).[265] While in large devices, a continuous switching between different polarization states is possible, devices scaled in the 10 nm regime show abrupt and accumulative switching.[259] The former can be used for mimicking synaptic functions, while the latter is helpful to mimic neurons. In classical non-volatile memories, the depolarization fields created by non-ferroelectric layers or portions of the layer in series to the ferroelectric are a concern for the retention of the device. However, when creating short and long-term plasticity in synaptic devices, this can be turned into an advantage such that the device retention can be tailored.

D. Spintronic materials for neuromorphic computing

Bernard Dieny and Tuo-Hung (Alex) Hou

1. Status

Spintronics is a merging of magnetism and electronics in which the spin of electrons is used to reveal new phenomena, which are implemented in devices with improved performances and/or new functionalities. Spintronics has already found many applications in magnetic field sensors, in particular in hard disk drives and, more recently, as non-volatile memory (MRAM) in replacement of e-FLASH and last-level cache memory. Spintronics can also bring very valuable solutions in the field of neuromorphic computing both as artificial synapses and as neurons.

Artificial synapses are devices supposed to store the potential weight of the bounds linking two neurons. Various types of spintronic synapses have been proposed and demonstrated.[65] They are magnetoresistive non-volatile memory cells working either as binary memory, as multilevel memory, or even in an analog fashion. Their resistance depends on the history of the current that has flown through the device (memristor). Most of these devices are based on magnetic tunnel junctions (MTJs), which basically consist of two magnetic layers separated by a tunnel barrier. One of the magnetic layers has a fixed magnetization (the reference layer), whereas the magnetization of the other (the storage layer) can be changed by either a pulse of magnetic field or current using phenomena such as spin transfer torque (STT) or spin–orbit torque (SOT).[266] The resistance of the device depends on the amplitude and orientation of the magnetic moment of the storage layer relative to that of the reference layer (tunnel magnetoresistance effect—TMR). For a binary memory as in STT-MRAM or SOT-MRAM, only the parallel and antiparallel magnetic configurations are used.[267] For multilevel or analog memory, several options are possible as illustrated in Fig. 13. One consists of varying the proportion of the storage layer area that is in parallel or antiparallel magnetic alignment with the reference layer magnetization. This can be achieved by step-by-step propagation of a domain wall within the storage layer using the STT produced by successive current pulses [Fig. 13(a)],[268] or by gradually switching the magnetization of the storage layer exchange coupled to an antiferromagnet using the SOT produced by the pulsed current flow in the antiferromagnet [Fig. 13(c)],[269] or by gradually switching the grains of a granular storage medium similar to the ones used in hard disk drives [Fig. 13(d)],[270] or by nucleating a controlled number of magnetic spin nanotextures in the storage layer, such as skyrmions [Fig. 13(e)].[271] Alternatively, the memristor resistance can also be varied by changing the relative angle between the magnetization of the reference and storage layers using all intermediate angles between 0° and 180° instead of only parallel and antiparallel configurations [Fig. 13(b)].[103] Chains of binary magnetic tunnel junctions can also be used to achieve spintronic memristors but at the expense of a larger footprint.[272]

Concerning artificial neurons, the conventional CMOS neuron circuit is limited by its large area because a large number of transistors and a large-area membrane capacitor are required for implementing Integrate-and-Fire (I&F) functions.[273] Recently, several spintronic neuron devices have been reported to generate spike signals by leveraging nonlinear and stochastic magnetic dynamics without the need for additional capacitors and complex peripheral circuitry. Spintronic neurons potentially show a great advantage for compact neuron implementation.[274]

Assembly of interacting spin-torque nano-oscillators (STNOs) based on the structure of magnetic tunnel junctions (MTJs) was proposed to achieve neuron functionality. An unstable conductance oscillation that mimics spike generation is induced at hundreds of MHz to several tens of GHz by flowing a current through the device. The frequency and amplitude of oscillation vary with the applied current and magnetic field. Torrejon *et al.* demonstrated spoken-digit or vowel recognition using such an array of nanoscale oscillators.[275]

Superparamagnetic tunnel junctions can also be used to mimic stochastic neurons. They have much lower thermal stability compared to the MTJ used for memory, so they stochastically switch between antiparallel (AP) and parallel (P) states due to thermal fluctuations, which is referred to as telegraphic switching.[276] This switching mode can be used to generate Poisson spike trains in spiking neural networks (SNNs) as well as for probabilistic computing.[277] A MTJ device with high thermal stability, which can implement not only synapses but also neurons in an all-spin neural network, was proposed by Wu *et al.*[278] The reduction in the thermal stability factor is induced by self-heating at a high bias voltage for neuron operations.[279] At a low bias, it stably stores weight information as synapses.

Many other new spintronic materials and mechanisms were also investigated for the feasibility of neuron devices, in particular based on magnetoelectric effects. For instance, by playing with magneto-ionic effects influencing the anisotropy at magnetic metal/oxide interfaces, the density of skyrmions[280] and even their chirality could be controlled electrically.[281] Jaiswal *et al.* designed a magnetoelectric neuron device for SNNs.[282] Zahedinejad *et al.* demonstrated that electrically manipulated spintronic mem-

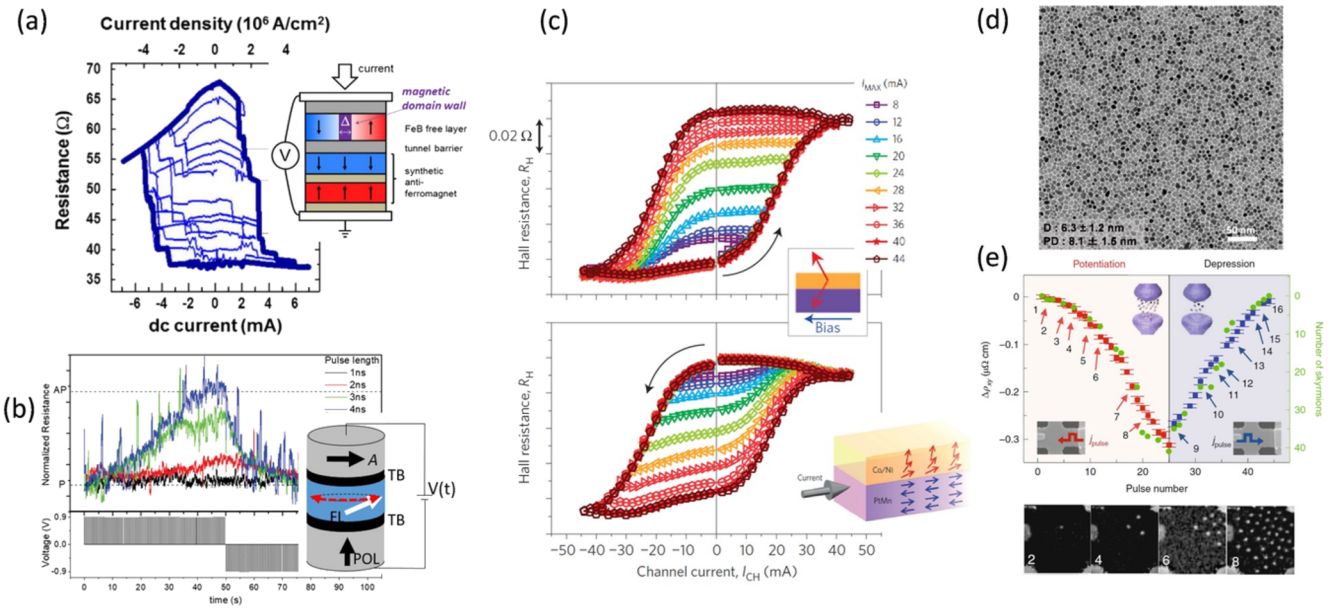

FIG. 13. Various realizations of spintronic memristors: (a) based on domain wall propagation in the storage layer; (b) based on variation of angle between storage layer and reference layer magnetization; (c) based on SOT in ferromagnetic storage layer exchange coupled to an antiferromagnetic SOT line; (d) implementing a storage layer made of a granular layer similar to the one used in recording technology; and (e) based on a controlled number of skyrmions nucleated in the storage layer.

ristors can be used to control the synchronization of spin Hall nano-oscillators for neuromorphic computing.[283]

2. Challenges

Building useful fully functional neuronal circuits requires large-scale integration of layers of artificial neurons interconnected with spintronic synapses. Crossbar architectures can achieve cumulate and multiply functions very efficiently in an analog manner. An advantage of magnetic tunnel junctions over other technologies based on materials such as resistive oxides or phase change is their write endurance associated with the fact that their resistance change does not involve ionic migration. However, they exhibit a lower R_{OFF}/R_{ON} ratio (~4 for MRAM vs 10–100 for RRAM or PCM) and also narrower cell-to-cell distribution of resistance in R_{OFF} and R_{ON} states. In crossbar architectures, MTJs should have high resistance to minimize power consumption. Therefore, efforts should be pursued to further increase the TMR amplitude of MgO-based MTJs and bring it closer to the expected theoretical values of several 1000%.[284] In high-resistance MTJs, other approaches, such as SOT or voltage control of anisotropy (VCMA), could be used to change the MTJ resistance. In all cases, the control of the resistance change induced by current or voltage pulses must be improved. The operating temperature has often also a significant impact on magnetic properties, which imposes challenges on system design.

Concerning artificial neurons, the DC power required to trigger the magnetization dynamics of STNO neurons is still relatively high (mW range).[285] Ways must be found to reduce it by using different materials or new designs. The switching speed and endurance in superparamagnetic tunnel junctions and self-heating-assisted MTJ neurons could be further enhanced to improve processing speed and system reliability.[286,287] How to continue improving variability across millions of synapses and thousands of neurons to ensure high accuracy in future neuromorphic systems remains an actively research topic. Interconnecting all these devices is also a challenge, and innovative approaches beyond classical interconnects must be found notably by taking advantage of 3D integration.

3. Potential solutions

STT-MRAM entered volume production in 2019 at major microelectronic companies.[288] This marked the adoption of this hybrid CMOS/magnetic technology by the microelectronic industry. Thanks to the combined efforts of the chip industry, equipment suppliers, and academic laboratories, spintronics is progressing very fast. Materials research is very important to increase magnetoresistance amplitude, switching currents, STT and SOT efficiency, and VCMA efficiency; reduce dependence on operating temperature; reduce current to trigger oscillations in STNOs; and reduce disturbance due to parasitic field. Investigations are in progress involving antiferromagnetic materials for reduced sensitivity to the field and access to THz frequency operation;[289] half-metallic materials, such as Heusler alloys, for enhanced TMR amplitude and reduced write current;[104] and topological insulators for very efficient spin/charge current interconversion possibly combined with ferroelectric materials.[290]

Concerning interconnects, fortunately, magnetic materials are grown in backend technology and can be stacked but at the expense of complexity and cost. Long-range information transmission can

be carried out via spin-current or magnons or by propagating magnetic textures, such as domain walls[291] or skyrmions.[292] Light could also be used to transmit information in conjunction with recent developments related to all-optical switching of magnetization.[293] In addition, a great advantage of spintronic stacks is that they can be grown on almost any kind of substrates, provided that the roughness of the substrate is low enough compared to the thickness of the layers comprised in the stack. This enables the use of the third dimension by stacking several spintronic structures, thereby gaining in interconnectivity.[294]

4. Concluding remarks

Spintronics can offer valuable solutions for neuromorphic computing. Considering that STT-MRAM is already in commercial production, it is very likely that the first generation of spintronic neuromorphic circuits will integrate this technology. Next, crossbar arrays implementing analog MTJs may be developed as well as neuronal circuits based on the dynamic properties of interacting STNOs for learning and inference. Still, many challenges are on the way toward practical applications, including speed, reliability, scalability, and variation tolerance, which need to be addressed in future research.

E. Optoelectronic and photonic implementations

Akhil Varri, Frank Brückerhoff-Plückelmann, and Wolfram Pernice

1. Status

Computing using light offers significant advantages in highly parallel operation exploiting concepts such as wavelength and time multiplexing. Moreover, optical data transfer enables low power consumption, better interconnectivity, and ultra-low latency. Already in the 1980s, first prototypes were developed; however, the bulky tabletop experiments could not keep pace with the flourishing CMOS industry. At present, novel fabrication processes and materials enable the (mass) production of photonic integrated circuits, allowing photonic systems to compete with their electronic counterparts. Especially in the area of data-heavy neuromorphic computing, the key advantages of photonic computing can be exploited.

Scientific efforts in neuromorphic photonic computing can be segregated in two major directions: (i) one approach is building hardware accelerators that excel at specific tasks, e.g., computing matrix–vector multiplications, by partially mimicking the working principles of the human brain, and (ii) the other is creating designs that aim to emulate the functionality of biological neural networks. Such devices are able to replicate the behavior of a neuron, a synapse, and learning mechanisms and to ultimately implement a spiking neural network.

There has been considerable progress in the (i) direction since 2017 when Shen *et al.*[295] demonstrated vowel recognition where every node of the artificial neural network is physically represented in the hardware using a cascaded array of interferometers. This scheme has also been scaled to implement a three-layer deep neural network with *in situ* training capability.[296] In addition, Feldmann *et al.*[297] have demonstrated neurosynaptic networks on-chip and used them to perform image recognition. The photonic circuit deploys a non-volatile phase change material (PCM) to emulate the synapses and exploits the switching dynamics as a nonlinear activation function. As highlighted in Sec. V B, the integration of PCMs also leads to in-memory computing functionality owing to their non-volatile nature.

For the (ii) direction, significant work has been done on a device level to mimic individual components of the brain. Excitable lasers combining different material platforms, such as III–V compounds, and graphene have been shown to demonstrate leaky integrate and fire-type characteristics of a neuron.[73] In addition, neurons based on optoelectronic modulators have been shown in the literature. For synapses, photonic devices combined with PCMs, amorphous oxide semiconductors, and 2D materials have been used to demonstrate synaptic behaviors, such as spike-time-dependent plasticity and memory.[73,298,299] Furthermore, key synaptic functions have also been demonstrated using optically controlled reversible tuning in amorphous oxide memristors.[300,301] Optical control of the conductance levels in memristors, such as those described in Sec. V A, enables low-power switching dynamics important to neuromorphic computing efforts.

In the following, we break down the challenge of building neuromorphic photonic hardware to various subtopics, ranging from increasing the fabrication tolerance of the photonic circuit to co-packaging the optics and electronics. Then, we review the current advances in those areas and provide an outlook on the future development of neuromorphic photonic hardware.

2. Challenges

A major challenge is combining the various building blocks shown in Fig. 14. Silicon on insulator is the platform of choice for building large circuits owing to the matured CMOS process flow and the high refractive index contrast between the silicon waveguide and oxide cladding. However, silicon has no second-order nonlinearity as it is centrosymmetric. Furthermore, silicon being an indirect bandgap material cannot emit light. This strongly limits the options for implementing nonlinear functions and spiking dynamics crucial for an all-optical neural network. Therefore, most of the research on mimicking neurons is focused on novel material platforms that support gain. A key challenge is integrating those different material platforms. For example, a circuit may deploy neurons based on III–V semiconductor heterojunctions and synapses built with PCMs on silicon. Therefore, compact and fabrication error-tolerant optical interconnects are crucial for the performance of the whole system.

Apart from packaging various optical components, the electro-optic interface imposes an additional challenge. Typically, the input data are provided by digital electrical signals, whereas optical data processing is analog. This requires analog-to-digital converters (ADCs) and digital-to-analog converters (DACs) for digital systems to interface with the chip as shown in Fig. 14. For large circuits, co-packaging electronics and photonics increases the footprint and cost significantly. This negatively affects the throughput.

In addition to the above, fabrication imperfections will also impact the performance of a photonic circuit. Components such as ring resonators, cavities, and interferometers employed in many photonic circuits are designed to operate at a certain wavelength.

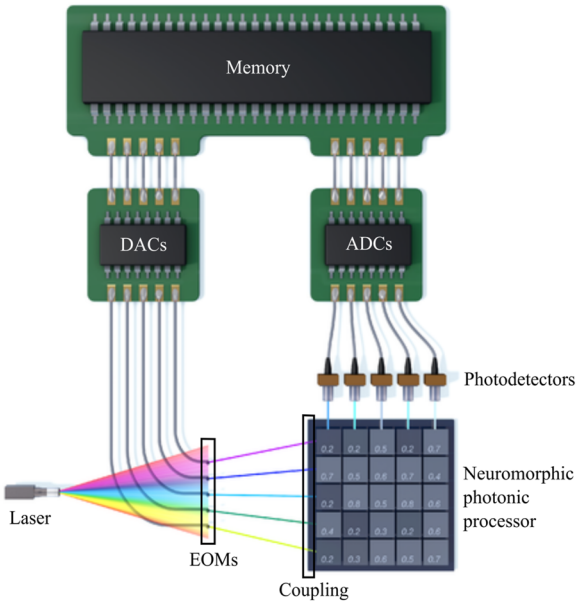

FIG. 14. An illustration of an optoelectronic system capable of performing computing operations. The electro-optic modulator (EOM) encodes the input data from the memory or a real-time sensor into the light fields. The encoded light fields are then coupled to a photonic processor that emulates the neurosynaptic behavior. The results from the processor are passed on to detectors that transform the signals back into the electronic domain and are finally stored in a memory.

However, due to factors such as etching rate, sidewall angle, and surface roughness, the wavelength of operation many times does not match with the design. Hence, in many cases, active methods such as thermo-optic phase shifters are employed to post-fabricate trim the wavelength. This results in unnecessarily increased electronic circuitry adversely affecting the scalability of the system.

Finally, a challenge that may be critical in the future is the electro-optic modulator (EOM) efficiencies that depend on the material properties and configurations. An important figure of merit for EOM efficiency is VpiL. This merit shows the voltage that needs to be applied and the length of the modulator required to obtain a pi phase shift to the input. A smaller merit figure suggests increased power efficiency and a compact footprint. As photonic neuromorphic circuits are supposed to scale up in the future, the power budget and space available on-chip will play an essential role in influencing the designs.

3. Potential solutions

Solutions addressing the challenges mentioned above lie on multiple fronts. First, we discuss how the scalability can be improved from a device-level perspective. The compact footprint, power efficiency, and cascadability of the neurons are essential characteristics for improving the scalability. In this regard, modulator-based neurons can be improved by integrating with materials such as electro-optic polymers. These materials have an order of magnitude higher r33 electro-optic coefficient compared to bulk lithium niobate, which has been conventionally the material of choice for modulators. As a result, electro-optic polymers integrated with silicon waveguides show very low VpiL among fast modulators.[302] In addition, novel materials, such as epsilon-near-zero (ENZ), which are promising for optical nonlinearity, can also be explored.[303] Nevertheless, for the widespread use of these devices, a better understanding of the material properties and engineering efforts to integrate them into the existing manufacturing process flow is required.

Particularly, integration techniques, such as micro-transfer printing, flip-chip bonding, and photonic wire bonding, will play a key role. To solve the problem of packaging with electronics, strategies such as monolithic fabrication, where the photonics and electronics are on the same die, need to be investigated. Foundries are now offering multi-project wafer runs with these state-of-the-art packaging techniques.

For improving the scalability of spike-based processing systems, another class of neurons that is very promising is the vertical cavity surface emitting lasers (VCSELs). VCSELs can integrate 100 picosecond-long pulses and fire an excitable spike when the sum crosses a certain threshold, emulating biological neurons. Recently, it has been shown that the output of one layer of VCSEL neurons combined with a software-implemented spiking neural network can perform 4-bit image recognition.[304] In order to build the entire system on hardware and perform larger experiments, 2D VCSEL arrays flip-chip bonded on a silicon die can be examined.

Finally, to address the challenge of fabrication imperfections, passive tuning approaches can be of interest, which need no additional circuitry and are non-volatile. One direction could be the use of phase change materials, such as GaS and Sb_2S_3, to correct for the variability in photonic circuits.[305] These materials are very interesting since their real part of the refractive index can be tuned while keeping low absorption at telecom wavelengths. Another approach for post-fabrication passive trimming could be to use an electron beam or ion beam to change the material properties of the waveguide. This method is also scalable as these tools are widely used in the semiconductor industry.

4. Concluding remarks

Applications such as neuromorphic computing are particularly promising for optics where their unique advantages (i.e., high throughput, low latency, and high power efficiency) can be utilized. At present, there have been instances in the literature where different devices have been proposed to emulate the individual characteristics of a neurosynaptic model. However, there is a lot of scope for research in materials science to pave the way for more compact, cascadable, and fabrication-friendly implementations. Furthermore, large-scale networks are expected to scale in the near future by integrating state-of-the-art packaging techniques that are now available to research groups and startups.

To summarize, the growth of integrated photonics has led to a resurgence of optical computing not only as a research direction but also commercially. It is exciting to see how the field of neuromorphic photonics will shape as advancements in science and technology continue to happen.

F. 2D materials

Mario Lanza, Xixiang Zhang, and Sebastian Pazos

1. Status

Multiple studies have claimed the observation of resistive switching (RS) in two-dimensional layered materials (2D-LMs), but very few of them reported excellent performance (i.e., high endurance and retention plus low switching energy, time, and voltage) in a reliable and trustable manner and in a device small enough to be attractive for high-integration-density applications (e.g., memory and computation).

The best RS performance observed in 2D-LMs is based on out-of-plane ionic movements. In such types of devices, the presence and quality of the RS phenomenon mainly depend on three factors: the density of native defects, the type of electrode used, and the volume of the dielectric (thickness and area). In general, 2D-LMs with excellent crystallographic structure (i.e., without native defects, such as those produced by mechanical exfoliation) do not exhibit stable resistive switching. Reference 306 reported that mechanically exfoliated multilayer MoS_2 does not show RS; only after oxidizing it (i.e., introducing defects), it shows RS based on the migration of oxygen ions. Along these lines, Ref. 307 showed that mechanically exfoliated multilayer hexagonal boron nitride (h-BN) does not exhibit RS; instead, the application of voltage produces a violent dielectric breakdown (DB) followed by material removal. The more violent DB phenomenon in h-BN compared to MoS_2 is related to the higher energy for intrinsic vacancies formation: >10 eV for boron vacancies in h-BN vs <3 eV for sulfur vacancies in MoS_2. Some articles claimed RS in mechanically exfoliated 2D-LMs, but very few cycles and poor performance were demonstrated; those observations are more typical of unstable DB than stable RS. Reference 308 reported good RS in a crossbar array of Au/h-BN/graphene/h-BN/Ag cells produced by mechanical exfoliation, but in that study, the graphene film shows an amorphous structure in the cross-sectional transmission electron microscope images. Hence, stable and high-quality RS based on ionic movement has never been demonstrated in as-prepared mechanically exfoliated 2D-LMs. This is something expected because ionic-movement-based RS is only observed in materials with high density of defects (e.g., high-k materials and sputtered SiO_2) but not in materials with low density of defects (e.g., thermal SiO_2), as the higher energy-to-breakdown forms an irreversible DB event.

On the contrary, 2D-LMs prepared by chemical vapor deposition (CVD) and liquid phase exfoliation (LPE) have exhibited stable RS in two-terminal memristors[309] and three-terminal (memtransistors) configurations,[310] although in the latter, the switching mechanism is largely different. In two-terminal devices, RS is enabled by the migration of ions across the 2D-LM. In transition metal dichalcogenides (TMDs), the movement of chalcogenide ions can be enough to leave behind a metallic path (often referred to as conductive nanofilament or CNF) that produces switching (similar to oxygen movement in metal-oxides).[306] However, in h-BN, metal penetration from the adjacent electrodes is needed, as this material contains no metallic atoms.[307] In 2D-LMs prepared by CVD and LPE methods, ionic movement takes place at lower energies (than in mechanically exfoliated ones) due to the presence of native defects (mainly lattice distortions and impurities). The best performance so far has been observed in CVD-grown ~6 nm-thick h-BN, as it is the only material with enough insulation and thickness to keep low the current in the high resistive state.[72] This includes the coexistence of bipolar and threshold regimes (the second one with highly controllable potentiation and relaxation), bipolar RS with endurances $>5 \times 10^6$ cycles (similar to commercial RRAM memories and phase-change memories),[311] and ultra-low switching energies of ~8.8 zJ in the threshold regime.[312] Moreover, a high yield (~98%) and low variability have been demonstrated.[312] In 2D-LMs produced by LPE or other solution-processing methods,[313] the junctions between the flakes and their size play a very important role, and while there is evidence of potentially good endurance, synaptic behavior, and variability, sub-μm downscaling still has not shown equivalent performance.[314]

Apart from ionic movement, 2D-LMs can also exhibit RS based on ferroelectric effect.[250] A remarkable example is In_2Se_3, which has electrically switchable out-of-plane and in-plane electric dipoles. Recent works have demonstrated that RS in ferroelectric In_2Se is ensured by three independent variables (polarization, initial Schottky barrier, and barrier change) and that it delivers multidirectional switching and photon storage.[315] However, the endurance and retention time are still limited to hundreds of cycles, and stable ferroelectric RS at the single-layer limit remains unexplored. Finally, tunable optoelectronic properties and unique electronic structure attainable through 2D-LM heterostructures present enormous potential for near-/in-sensor computing in neuromorphic systems. The responsiveness to physical variables (light, humidity, temperature, pressure, and torsion) of 2D-LM memristor and memtransistor devices allows us to mimic biological neurosynaptic cells (visual cortex and tactile receptors).[316]

2. Challenges and potential solutions

The main challenge of RS devices (of any type) is to exhibit high endurance in small devices. Many studies have reported RS in large devices with sizes >10 μm^2 and claimed that their devices are "*promising*" for memory and computing applications. This is a huge and unreasonable exaggeration; these two applications require high integration density, as commercial devices for those applications have sizes down to tens or hundreds of nanometers. It should be noted that in ionic-movement-based RS devices, the CNF always forms at the weakest location of the sample; when the device size is reduced, the density and size of defects are (statistically) reduced, which produces an increase in the forming voltage.[317] Hence, the CNF of smaller devices is wider due to the larger amount of energy delivered during the forming. This has a huge effect on state resistances, switching voltages, time, and energy, as well as endurance, retention time, and device-to-device variability. In other words, the fact that a large device (>10 μm^2) exhibits good RS does not mean that a small device (<1 μm^2) made with the same materials will also exhibit it; hence, RS "*promising for memory and computation*" is only the one that is observed in devices with sizes of tens/hundreds of square nanometers.

Taking this into account, the main challenge in 2D-LM based devices is to observe RS in small devices, and the most difficult figure-of-merit to obtain is (by far) the endurance. Reference 318

demonstrated good RS in 5 × 5 μm^2 Au/h-BN/Au devices, in which h-BN was ~6 nm-thick and grown by the CVD method; however, when the size of the devices was reduced to 320 × 420 nm^2, the yield and the number of devices observed were very limited. The main issue was the current overshoot during the switching, which takes place randomly and produced irreversible DB in most devices. Similarly, solution-processed Pt/MoS_2/Ti devices[319] showed excellent performance across all figures-of-merit observed in 25 μm^2 devices, but such performance has not been reported for 500 × 500 nm^2 devices patterned via electron beam lithography. In this case, the large size of the nanoflakes (slightly below 1 μm minimum) may be imposing an intrinsic scaling limitation. Meanwhile, the scaling and overshoot problem was problem was solved in Ref. 72 integrating CVD-grown h-BN right on top (via the wet-transfer method) of a silicon complementary metal–oxide–semiconductor (CMOS) transistor, which acted as an instantaneous current limitation. Moreover, this approach brings the advantage of a very small device size (in Ref. 72, it was only 0.053 μm^2, as the bottom electrode of the RS device is from one of the metallization levels). The heterogeneous integration of the 2D-LMs at the back-end-of-line (BEOL) wiring of silicon microchips could be a good way of testing materials for RS applications and directly integrating selector devices with each memristor (into one transistor–one memristor (1T1M) cells), which is fundamental for the realization of large memristive synapse arrays—all state-of-the-art demonstrations of memristive neural accelerators based on mature memristor devices use 1T1M cells or differential implementations of such cells (2T2M and 4T4M). So far, these CMOS testing vehicles for RS materials have mainly been employed by the industry; in the future, spreading this type of testing vehicles among academics working in the field of RS could improve the quality of the knowledge generated. In addition, these devices may benefit from common practices in the field of silicon microchip manufacturing, such as surface planarization, plug deposition, and high-quality thick interconnect techniques.

Next steps in the field of 2D-LMs for RS applications consist in improving the materials quality to achieve better reproducibility of the experiments (from one batch to another) and adjust the thickness and density of defects to achieve better figures-of-merit in nanosized RS devices while growing 2D-LM at the wafer-scale.[320] Recent studies successfully synthesized large-area single-crystal 2D-LMs via CVD,[321] although in most cases, it is only monolayer. However, monolayer 2D-LMs are less than 1-nm-thick, and when they are exposed to an out-of-plane electrical field, a very high leakage current is generated even if no defects are present, which increases a lot the current in the high-resistance state (HRS) and the energy consumption of the device. Reference 322 presented the synthesis of single-crystal multilayer h-BN using scalable methods, but controlling the number of layers is still difficult. Electrical studies in such types of single-crystal multilayer samples should be conducted. Improving manipulating methods to prevent the formation of cracks during transfer is also necessary, although it is worth mentioning that multilayer h-BN materials are more mechanically stable than monolayers.

Recent demonstration of vector–matrix multiplication using MoS_2 memtransistors[323] is a promising advance in terms of a higher-level functional demonstration, although the fundamental phenomenon exploited is the well-known floating gate memory effect, not unique to 2D-LM themselves. Meanwhile, understanding the role of flake size in the functionality of solution-processed 2D-LM two-terminal synaptic devices is critical to address the true scaling limitations of such an approach, a key aspect to define potential realistic applications in neuromorphic systems. On the other hand, sensing capabilities emerge with great potential for biological synaptic mimicking. The full potential of different 2D-LM material heterostructures and memtransistors opens a huge design-space worth of exploration. In that sense, the complex physical characteristics offered by different 2D-LMs hold the potential not only for basic neuromorphic functionality but also for higher-order complexity. This could be exploited to achieve high-complexity neural and synaptic functions,[100] more closely mimicking actual biological systems. However, in parallel to elucidating the physical properties and capabilities of these material systems, efforts should be put into strengthening the quality of the reported results, focusing on proper characterization methods, reliable practices, and statistical validation.

3. Concluding remarks

Leading companies, such as TSMC, Samsung, IBM, and Imec, have started to work with 2D-LMs, but mainly for sensors and transistors. In the field of 2D-LM based neuromorphic devices, most work is being carried out by academics. In this regard, unfortunately, many studies make a simple proof-of-concept using a novel nanomaterial without measuring essential figures-of-merit, such as endurance, retention, and switching time. What is even worse, in many cases, the studies employ unsuitable characterization protocols that heavily overestimate the performance (the most popular case is the erroneous measurement of endurance[324]), withholding information regarding the failure mechanisms that lead to certain performance metrics not being achieved on some devices. This working style often result in articles with striking numbers (i.e., performance), but those are unreliable, and it is really bad for the field because it creates a hype of expectations and disillusion among investors and companies. The most important is that the scientists working in this field follow a few considerations: (i) always aim to show high performance in small (<1 μm^2) devices fabricated using scalable methods (even better if they are integrated on a functional CMOS microchip, not on an unfunctional SiO_2 substrate); (ii) measure all the figures-of-merit of several (>100) memristive devices for the targeted application (this may vary depending on the application);[214] (iii) clearly define the yield-pass criteria and the yield achieved, as well as the device-to-device variability observed; and (iv) whenever a failure mode is observed preventing reaching a desired figure-of-merit, clearly convey it to maximize the probabilities of finding a solution.

VI. MATERIALS CHALLENGES AND PERSPECTIVES

Stefan Wiefels and Regina Dittmann

A. Materials challenges

For the neuromorphic computing approaches addressed in Sec. III, the use of emerging memories based on novel materials will

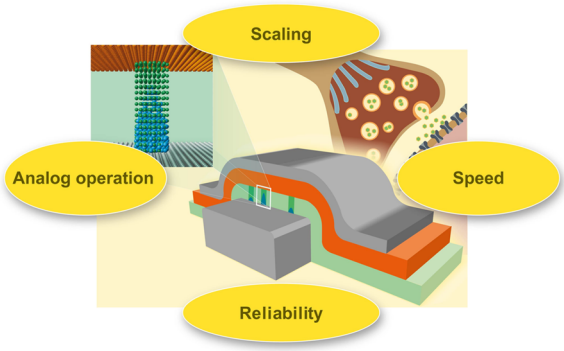

FIG. 15. Materials challenges for neuromorphic computing. Novel NVMs need to be scalable, fast, and reliable and allow for analog operation.

be key in order to improve their performance and energy-efficiency. This chapter discusses the most relevant properties and challenges for different use cases and how they relate to the respective materials properties. However, it is important to note that a dedicated co-development of materials with the readout and write algorithms and circuitry will be required in order to advance the field (Fig. 15).

1. Scaling

One main driving force to use emerging materials and devices is to gain space and energy efficiency by the fabrication of highly dense crossbar arrays. For STT-MRAM, scaling down to 11 nm cells has been demonstrated, as well as the realization of 2 Mb embedded MRAM in 14 nm FinFET CMOS.[325] However, due to the small resistance ratio of 2–3, the readout of magnetic tunnel junctions (MTJs) is more complex than for other technologies. Nevertheless, a 64 × 64 MTJ array, integrated into 28 nm CMOS, has recently been realized.[75] Advancements from the material side will be needed in order to increase the resistance ratio of MTJs in the future.

For ferroelectric HfO_2-based devices, the main challenge with respect to scaling is to decrease the thickness reliably in order to enable 3D capacitors with 10 nm node and to obtain a uniform polarization at the nanoscale of a material that currently still contains a mixture of different phases. Therefore, ultrathin films with the pure ferroelectric orthorhombic phase and without any dead-layers at the interfaces will be key to approach the sub-20 nm regime of hafnia-based ferroelectric devices.[9]

PCM devices can be fabricated on the sub-10 nm scale.[326] The limiting factor for CMOS integrated PCM devices is the high RESET current, which is required to implement larger access transistors.[202] Commercially available ReRAM cells with conventional geometries have been co-integrated on 28 nm CMOS technology. By employing a sidewall technique and nanofin Pt electrodes, small arrays with 1×3 nm^2 HfO_2[202] cells and 3×3 arrays of $Pt/HfO_2/TiO_x/Pt$ cells with a 2 nm feature size and a 6 nm half-pitch have been fabricated, respectively.[327]

With respect to ultimate scaling, the loss of oxygen to the environment might pose limitations to the retention times for ReRAM devices scaled in the sub-10 nm regime.[328] However, filaments in the size of 1–2 nm can be stable if they are stabilized by structural defects, such as grain boundaries or dislocations. Therefore, finding materials solution for confining oxygen vacancies to the nanoscale might retain the required retention for devices in the few nm scale.

2. Speed

Although the extensive parallelism leads to high demands for scaling, it is considered an advantage as it makes the race for ever increasing clock frequencies obsolete.[329] In contrast, the operation speed is closely linked to the respective application, i.e., the timing is based on real physical time.[329] As signals processed by humans are typically on a time scale of milliseconds or longer, the expected speed benchmark is well below the reported speed limits of emerging NVMs. Nevertheless, it is reasonable to understand the ultimate speed limits of NVM concepts in order to estimate maximum learning rates, to explore the impact of short spiking stimulation. Furthermore, novel computing concepts, as discussed in Sec. VII, might still benefit from higher clock frequencies. For MRAM, reliable 250 ps switching has been demonstrated by using double spin-torque MTJs, which consist of two reference layers, a tunnel barrier, and a non-magnetic spacer.[330] FeRAM arrays have successfully been switched with 14 ns at 2.5 V. Ferroelectric field effect transistors (FeFETs) have been shown to switch with <50 ns pulses in 1 Mbit memory arrays.[9] PCM devices can be switched with pulses <10 ns.[326] In general, their speed is limited by the crystallization time of the material. It has been shown exemplarily on Ge_xSn_yTe samples that this time can be tuned in a broad range of 25 ns up to 10 ms by adjusting the material composition.[331] Thus, it has a high potential to match the operation time of an NC system to the respective application. For VCM ReRAM, SET and RESET switching with 50 and 400 ps have been demonstrated.[332] Both are so far limited by extrinsic effects and device failure modes rather than by intrinsic physical rate limiting steps.

3. Reliability

Independent of the application, the reliability of the memory technology has to be taken into account. In the case of implementing NVMs as artificial synapses, the requirements of learning and inference phases have to be distinguished. Whereas the endurance is more relevant for learning schemes, the stability of the programmed state, i.e., the retention and robustness against read disturb, has to be sufficient for reliable inference operations.

4. Endurance

While MRAM has, in principle, unlimited endurance, all memristive devices that are based on the motion or displacement of atoms, such as ReRAM, PCM, and ferroelectric systems, have limited endurance. For silicon-based FeFETs, the endurance is typically on the order of 10^5, which is mainly limited by a dielectric breakdown in the SiO_2 at the $Si–HfO_2$ interface.[9] Regarding the endurance of VCM ReRAM, it has been demonstrated with convincing statistics that $> 10^6$ cycles are realistic. Some reports suggest maximum cycle numbers of more than 10^{10}.[324] Depending on the material system, various failure mechanisms for endurance are discussed. The microstructure of the switching material might degrade or be irreversibly penetrated by metallic atoms.[9] In VCM ReRAM, an excessive generation of oxygen vacancies was discussed as an endurance limiting factor.[333] Novel material solutions, which confine ions to the intended radius of action, might be a pathway to increase the endurance of

ReRAM devices. For PCM, it was suggested to implement multi-PCM synapses. Arbitration over multiple memory elements might circumvent endurance and variability issues.[9]

A typical limitation with respect to a reliable operation of ferroelectric memories is the so-called wake-up effect, which causes an increasing polarization after a few cycles and the fatigue resulting in a decrease in the polarization for high cycle numbers. Both are induced by the motion of defects such as oxygen vacancies and will have to be tackled in the future by intense materials research in this field.

5. Retention

After training, the state of the non-volatile memory synapse is required to be stable for 10 years at an operating temperature of 85 °C. However, for many applications in the field of neuromorphic applications, the requirement is much more relaxed in particular for the training phase. From a thermodynamical point of view, the states in ferroelectric or ferromagnetic memories might both be stable. In contrast, ReRAM and PCM devices store information in the configuration of atoms where both low-resistance state (LRS) and HRS are metastable states and the retention is determined by material parameters, such as the diffusion coefficient of the respective species.[9] Here, the degradation is not a digital flipping of states but a gradual process. For PCM, the drift of the resistance state is caused by the structural relaxation of the melt-quenched amorphous phase.[202] Apart from a drifting of the state, a broadening of the programmed state distribution (e.g., resistance) is typically observed for ReRAM.[334] Furthermore, since analog or multi-level programming is highly relevant for NC, it should be considered that intermediate resistance states might have a reduced retention compared to the edge cases of high and low resistive states as demonstrated for PCM devices.[9]

6. Read disturb

During inference, frequent reading of the memory elements is required, which should not change the learned state. For a bipolar ReRAM memory, a read disturb in the HRS/LRS occurs mainly when reading with a SET/RESET polarity since the read-disturb can be considered as an extrapolation of the SET/RESET kinetics to lower voltages. Nevertheless, the HRS state in bipolar filamentary VCM has been demonstrated by extrapolation to be stable for years at read voltages up to 350 mV.[335]

7. Variability

Variability is most pronounced for systems that rely on the stochastic motion and redistribution of atoms, such as ReRAM and PCM. Here, the variability from device to device (D2D), from cycle to cycle (C2C), and even from one read to the next (R2R) has to be distinguished. By optimizing fabrication processes, the D2D variability can be kept comparatively low. In contrast, the C2C variability for filamentary resistive ReRAM and PCM can be significant due to the randomness of filament[335] or crystal[202] growth, respectively. However, using smart programming algorithms, the C2C variability can be very well reduced to a minimum.[335] By contrast, R2R variations remain in the form of read noise in filamentary VCM. It is typically attributed to the activation and deactivation of traps or the random redistribution of defects[335] and strongly depends on the material.[336] For PCM, R2R variations are caused by 1/f noise and temperature induced resistance variations. One approach to address these issues as well as the drift is to use the so-called projected phase change memory with a non-insulating projection segment in parallel to the PCM segment.

Although the variability is a challenge for storage applications, it might be possible to design neuromorphic systems to exploit it.[329] In the end, a thorough understanding of the intrinsic variability might enable to match neuromorphic applications and materials.[336]

8. Analog operation

For most computing concepts described in Sec. III, their operation with binary memory devices is strongly limited and the possibility to adjust multiple states is of crucial importance. For devices with thermodynamically stable states, such as ferroelectric or magnetic memory, intermediate states rely on the presence of domains. As a result, the performance strongly depends on the specific domain structure and scaling might be limited by the size of the domains. Nevertheless, multilevel switching has been demonstrated by fine-tuning of programming voltages for both FTJ and FeRAM.[337]

For ReRAM and PCM, the metastable intermediate states have to be programmed in a reliable manner. Since these states are kinetically stabilized during programming, the success depends strongly on the switching kinetics of the specific system, the operation regime, and the intrinsic R2R variability of the material. For PCM devices, intermediate states can be addressed by partial reset pulses, which result in partial amorphization. As a result of the crystallization kinetics, a gradual crystallization can be obtained by consecutive pulses.

Filamentary ReRAM devices usually undergo an abrupt SET, which is caused by the self-accelerating, thermally driven filament formation. However, it is possible to obtain intermediate states with good control of the SET current, by precisely controlling the timing of the SET voltage pulses, or by slowing down the switching kinetics. This is the case for non-filamentary systems, which show a very pronounced gradual behavior for both SET and RESET.[327]

Furthermore, resistive switching devices with purely electronic switching mechanisms, such as trapping and de-trapping of electrons at defect states, might be promising for analog operation.[338]

B. Characterization techniques

Adnan Mehonic, Wing H. Ng, Mark Buckwell, Horatio R. J. Cox, Daniel J. Mannion, and Anthony J. Kenyon

1. Status

Memristive devices pose challenges to the experimentalist both in investigating the physics underpinning the device behavior and in optimizing functionality. The wide range of physical phenomena involved in memristance—from metal diffusion in dielectrics to Mott metal–insulator transitions, phase changes, and the formation of oxygen vacancy filaments—requires comprehensive physical, electrical, and chemical characterization and even the development

of novel analytical techniques.[339] Here, rather than an exhaustive literature survey, we provide examples to illustrate the recent progress in characterization of memristive materials and devices. While we concentrate on oxide-based RRAM materials and devices for reasons of space, most techniques reviewed here are applicable to other memristor types (PCM, MRAM, and FeRAM), which require similar characterization of structural, chemical, and electrical changes occurring in devices as a result of operation. The challenges presented to the experimentalist, particularly when it comes to structural and chemical analyses, are largely similar across all memristive devices.

2. Challenges and potential solutions

Looking first at resistance switching materials and devices, we see that early work from the 1960s on dielectric breakdown suggested the formation of conductive filaments in electrically biased oxide[340] and, while experimental techniques at the time were unable to image them, the authors correctly surmised their existence. The small sizes of these filaments, which can be of the order of a few nanometers in diameter, make studying their formation and disruption difficult. This is particularly true for oxygen vacancy filaments in oxides; the minimal contrast between the oxide matrix and the oxygen-deficient filament when imaged using electron beam techniques, such as TEM-EELS, means that there are few direct observations of such filaments. This contrasts with several published TEM studies of metal filaments in oxides, including the seminal work by Yang *et al.*,[341] which demonstrated field-driven movement of silver ions through SiO_2, Al_2O_3, and amorphous silicon to form dendritic conductive filaments. The large contrast between metal ions and oxide matrix makes it a more tractable problem to image individual conductive filaments. Subsequent work demonstrated different filament growth modes that depend on the relative magnitudes of metal ion mobility and the applied field.[342] The work by Waser *et al.*[343] details earlier TEM work on electrical and physical characterization of resistive switching dielectrics, including observations dating from as early as 1976 of silver dendrites formed in AgS under the application of an external field. Here again, the contrast between the silver filaments and the surrounding matrix provides a significant advantage. It is worth noting that the use of TEM measurements to characterize phase change memory (PCM) materials and devices is rather easier than in the case of oxide-based RRAM as the contrast between amorphous and crystalline phases of PCM materials is easier to detect, although *in situ* and in operando measurements can pose significant challenges, thanks to the fast switching speeds of PCM devices.[344]

The difficulty of imaging oxygen vacancy filaments using electron beam techniques can be overcome using conductive atomic force microscopy (CAFM) tomography ("scalpel AFM"), a review of which can be found in Ref. 345. In this technique, sequential CAFM scans of a sample surface imaged using a conductive diamond tip contacting the sample with sufficient force to scrape away the surface provide layer-by-layer conductivity maps that, when stacked, provide three-dimensional images of conductive regions within the oxide. Such studies reveal that electroforming generates large-scale changes beneath device top electrodes, modifying the conductivity of large volumes of material, while one or more highly localized conductive filaments bridge the inter-electrode gap. The technique also reveals details of the internal microstructure of the oxide, showing, for example, the columnar structure of sputter-deposited oxides, as the edges of columns are more conductive than their cores.[346] It should be noted that this technique only maps conductive regions that are connected to the bottom electrode, so they may be better thought of as a measure of *connectivity* rather than of regions of high conductivity.

The need to apply multiple analysis techniques to probe the movement of oxygen within, and emission of oxygen from, oxides under electrical stress was demonstrated by Mehonic *et al.*[347] TEM, EELS, CAFM, XPS, and mass spectrometry measurements of oxygen emission from samples under electrical stress combine with atomistic modeling to give a fuller picture of the dynamics of oxygen movement and its role in resistance switching. It is clear from such measurements that electrically biasing oxides—particularly those with some structural inhomogeneities—can drive large-scale changes in stoichiometry, reinforcing the CAFM tomography results referred to above. It has been known for some time that this can cause surface distortions and localized bubbling of both electrode and oxide surfaces.[340,348] While the work reported in Ref. 347 examines such features using AFM and TEM, the question of how mobile oxygen interacts with electrode materials is partly addressed in a recent work by Cox *et al.*[349] Oxide-based RRAM devices rely on the repeated reduction and oxidation of the switching oxide. To compete with high density flash, at least 10^4 cycles are required, and for many applications, more than 10^7 are needed. For such high numbers of cycles, oxygen should not be lost from the switching region around the conductive filament, implying the need for an oxygen reservoir that can both accommodate and release oxygen under appropriate electrical biases. This may be within the oxide, at an oxide/electrode interface, or within one or the other electrode. The need to measure oxygen movement is critical. Cox *et al.*[349] demonstrated that both the electrode material and the microstructure of the oxide layer influence the reversibility of oxygen movement. In the case of electrode metals with high oxygen affinities, oxygen moves easily from the oxide into the electrode, but its movement back again varies significantly between metals. Molybdenum, for example, both accepts and releases oxygen readily when the bias polarity is reversed, while titanium is easily oxidized when positively biased (and when neutral) but does not release oxygen back again when negatively biased. Platinum, having a very low electron affinity, does not accept or release oxygen.

Oxide porosity and sensitivity to moisture are important factors in resistance switching both by metal diffusion (extrinsic switching) and oxygen vacancy formation (intrinsic switching).[350] The presence of moisture in the switching oxide can lead to a highly variable resistance switching behavior. While the origin of such effects remains somewhat unclear, electrical measurements must be interpreted with care in the presence of moisture. The difficulty in measuring hydrogen content in materials and devices reliably and quantitatively is a particular challenge, not only for RRAM but also for other memristive devices. Hydrogen, even in relatively low concentrations, can affect the electrical properties of oxides and electrode stacks, as has been recognized for decades in the CMOS community. However, there have been very few studies on the role of hydrogen in memristive devices. At the same time, there is considerable interest in reducing variability and increasing stability in memristors. One cannot help but suppose that detailed

studies of hydrogen, and control of its presence, could help these efforts.

Electrical characterization of memristive devices poses other challenges. A critical issue in RRAM is electroforming conductive filaments. While some materials and devices are forming-free, the majority require an initial conditioning step in which a voltage higher than the normal programming (SET or RESET) voltages is applied to form a conductive filament that will subsequently be partially oxidized (RESET) and reduced (SET) by a sequence of voltage sweeps or pulses. Electroforming causes an abrupt drop in oxide resistance, which can span several orders of magnitude, over a timescale of nanoseconds or shorter. Without an appropriate limit on delivered current, this can destroy the device by irreversible oxide breakdown due to Joule heating. It is, therefore, essential to implement fast current limiting during electroforming. Response times of standard characterization instruments are generally too long, so current overshoots can over-stress the oxide, leading to breakdown or to conductive filaments too large (hence, too strong) to reset. Consequently, full electrical characterization of devices requires integrated current limiting devices, such as transistors or series resistors. Care must be taken to avoid parasitic capacitances or inductances, so the most reliable methods involve on-chip series resistors or transistors in the 1R1R or 1T1R configurations, where the first R or T label refers to the integrated current limiter (resistor or transistor) and the second R refers to the resistance switching element.

On the other hand, electrical characterization of RRAM devices can reveal important clues to the physical mechanisms responsible for resistance switching. Careful analysis of current/voltage curves can indicate a range of electron transport mechanisms, including various forms of tunneling (direct, trap-assisted, and Fowler–Nordheim), thermally assisted transport (Poole–Frenkel), and Ohmic conduction. These, in turn, provide evidence for microscopic processes, such as charge trapping/detrapping, formation of Schottky barriers, or various interface-related electronic states. However, more than one transport mechanism may contribute as, for example, currents may flow in parallel both through a conductive filament and through a highly defective surrounding oxide.[351] In the case of nanoscale filament formation, the thinnest point of the filament can behave as a quantum constriction, allowing only currents that are multiples of the conductance quantum, G_0, to flow. Such effects can be seen easily at room temperature.[351] Reviews of electrical characterization techniques for resistance switching devices can be found in Refs. 324 and 352.

3. Concluding remarks

The inherently interdisciplinary approach that is needed to fully characterize memristive devices poses challenges to the experimentalist. A wide range of techniques are required, which is often beyond the capabilities of a single laboratory, and in some cases, these techniques are operating close to their limits. While there have been significant advances in characterization (CAFM tomography, for example), more work is needed, for instance, to better characterize the role of hydrogen in resistance switching. Where characterization has been most successful has been when there is close collaboration, not only between experts in different experimental techniques but also with theorists who provide models to interpret experimental results.

C. Comparison between different material systems

Yuchao Yang and Yingming Lu

1. Status

Various types of memristors can be realized based on the abundant resistive switching mechanisms in different materials. Each type of memristor has certain characteristics (such as power consumption and switching speed) that are suited for specific applications, while different materials also have their own shortcomings and limitations. A clear understanding of the respective advantages and shortcomings of each material is key to its development and application.

Among various types of resistive switching materials, transition metal oxides (TMOs) are the most widely used due to their rich resistive switching mechanisms and characteristics. The retention time of TMO-based memristive devices, which indicates how long the resistive state can be maintained after electrical stimulation, is distributed in a wide range from μs to years. According to the retention time, TMOs can be generally divided into non-volatile TMOs and Mott TMOs.

The resistive switching of non-volatile TMOs, such as HfO_x and TaO_x, originates from the migration and redox reactions of oxygen ions or vacancies driven by an external electric field or thermal effects, which in turn create or destroy conductive filaments between the electrodes. The filaments can exist stably for a long time, and the continuous electrical modulation of geometric characteristics of the filaments, such as their lengths or diameters, results in multilevel resistive states. Therefore, the non-volatile TMO can be used to imitate the long-term plasticity (LTP) of biological synapses[169] and can accelerate the computationally intensive matrix–vector multiplication (MVM) in artificial neural networks. Furthermore, due to the mature and CMOS-compatible manufacturing process, a variety of in-memory computing chips based on non-volatile TMOs have been demonstrated.[56] One of the challenges of non-volatile TMO in circuit applications is the existence of the forming process, which requires a high voltage to initialize the TMO layer and subsequently increases the requirements for the voltage and robustness of peripheral circuits. In addition, due to the switching mechanism of non-volatile TMOs, the conductance change during the programming process usually shows nonlinear characteristics, along with obvious variations and noises, which increases the difficulty of programming, for example by necessitating closed-loop write-and-verify programming.

There are many other similar TMO systems, such as WO_x and TiO_x. Depending on the robustness of the formed filaments, the retention time can be gradually reduced as a result of filament dissolution, and the conductance of volatile TMOs can undergo a continuous decay process. This can effectively map information in time series into high-dimensional vectors, which is widely used in reservoir[184] for image classification or time series prediction.

Mott TMOs, mainly including VO_2 and NbO_2, show high resistance in body-centered tetragonal (BCT) or monoclinic (M) phases at low temperatures. Once the temperature of Mott TMOs exceeds a threshold, the reversible Mott transition occurs and results in

structural transition into the rutile (R) phase, accompanied by a significant resistance drop. Based on threshold switching, Mott TMOs are widely used in artificial neurons for constructing neuromorphic computing or sensory systems.[353] Meanwhile, the nonlinear transport mechanism of NbO_2 shows high-order complexity during current sweeps, which can effectively realize the rich dynamics of biological neural systems.[7] However, as the Mott transition is closely related to temperature, the operation of Mott TMOs can be affected by ambient temperature. On the other hand, this could serve as the physical foundation for temperature sensing. Furthermore, the metallic domains produced by previous switching events have been found to remain in VO_2 for a long time, which will affect the switching threshold voltage of the devices[354] subsequently. The transition temperatures of VO_2 and NbO_2 are around 70 and 800 °C, respectively, which are too low and too high from the perspective of circuit applications. A Mott TMO with ideal transition temperature is yet to be developed.

Phase change materials mainly refer to chalcogenide glass materials with reversible phase transition processes. Phase change materials show high resistance in the amorphous state and relatively low resistance in the crystalline state. The phase change memory (PCM) constructed by these materials shows multilevel non-volatile conductance states, and hence, PCM can also be exploited to accelerate MVM combined with a crossbar.[355] Traditional phase change materials are the ternary GeSbTe compounds along the pseudo-binary tie lines of GeTe–Sb_2Te_3, Ge–Sb_2Te_3, and GeTe–Sb.[356] By adjusting the proportions of the three elements in the compound or by incorporating other elements into the compound, these phase change materials can exhibit advantages in various aspects, such as ON/OFF ratio, switching speed, and retention. Among these materials, $Ge_2Sb_2Te_5$, with its outstanding recyclability, is currently widely used and has been applied in commercial products. In addition to compound-type phase change materials, monatomic Sb exhibits completely different resistances in the crystalline and amorphous states, making it also suitable as a phase change material.[357] This type of simplest material can effectively avoid the problem of stoichiometry deviation during the phase change process, which is important for further reducing the size of PCM devices. One of the challenges faced by PCM comes from its conductance drift, caused by the spontaneous structural relaxation in unstable amorphous materials, which leads to a gradual conductance decay over time and seriously affects computing accuracy and reliability. Therefore, a compensation circuit or strategy for conductance drift is desired. Another challenge originates from the long heating time required for the crystallization of PCM, which not only affects the programming speed[358] but also results in higher energy consumption during programming.

In addition, magnetic material-based magnetic random access memory (MRAM) is also widely used in neuromorphic computing, which has relatively mature technology, high endurance, etc. The typical structure of MRAM is a magnetic tunnel junction (MTJ), which is composed of two layers (pinned and free) of a magnetic material and an insulator (usually MgO) sandwiched in between. There are two types of magnetic materials mainly used in MTJs.[359] The first one is the multilayers formed by transition metals (e.g., Co and Fe) and noble metals (e.g., Pt and Pd), such as $(Co/Pd)_n$ and $(Co/Pt)_n$. These materials have advantages in terms of thermal stability and scalability. Another important magnetic material is CoFeB, which shows extremely low programming currents and good matching with the lattice of the MgO barrier layer. By tuning the magnetization orientations of the free magnetic layer to parallel (P) or antiparallel (AP) orientation with the pinned layer, MTJs can exhibit low or high resistance, respectively. Because only the magnetization orientation of the material is changed without large-scale atomic migration or rearrangement, MRAM has low device variation and high reliability. However, the main shortcoming of MRAM is reflected in its relatively low resistance even in the high resistance state,[75] which will increase the power consumption for MVM computing. Moreover, it has a small ON/OFF ratio, which supports only two states, indicated P and AP, and limits its applications in analog computing.

Similar to MTJs, ferroelectric tunneling junctions (FTJs) also change their resistance by adjusting the polarization orientations of the ferroelectric material sandwiched between two metal electrodes. The commonly used ferroelectric material systems include perovskite oxides, fluorite ferroelectric materials, and wurtzite ferroelectric materials. Perovskite oxides are the most widely used ferroelectric materials with advantages of scalability and switching speed.[360] Among these materials, $Pb[Zr_xTi_{1-x}O_3]$ (PZT) and $Sr_2Bi_2TaO_9$ (SBT) have more mature technology and have been applied in commercial memory devices. The fluorite ferroelectric materials, such as $HfZrO_2$, are also widely utilized and studied due to their CMOS-compatible fabrication process and good scalability.[361] The wurtzite materials, such as $Al_{1-x}Sc_xN$, belong to a new type of ferroelectric material system, characterized by their high ON/OFF ratio, thermal stability, and retention.[362] Since the polarization orientations of the ferroelectric layer can be tuned by the electric field with a very low tunneling current, the programming of FTJs consumes low power. Meanwhile, the FTJ shows multilevel conductance[363] for MVM acceleration in the edge AI platforms. However, ferroelectric materials can have retention degradation caused by intrinsic depolarization fields in the ferroelectric layer, while recent Hafnium Zirconium Oxide (HZO) based ferroelectric materials exhibit relatively high coercive voltages and low endurance.

Ion-gated transistors (IGTs) are a type of three-terminal device, where the electric field from the gate drives small ions (such as H^+ and Li^+) in electrolytes into the device channels to continuously tune the conductance of the channel.[364] Among the reported material systems for IGTs, the most chosen material for channel is TMOs, which shows the potential for mass production and environmental stability. For the material consideration of electrolyte, the phosphosilicate glass shows clear advantages over previously used Li-ions electrolyte, which is more mature and compatible in the COMS process platform. Owing to the low gate leakage current and separated programming and reading terminals, the IGT shows significantly lower power consumption during programming compared with most two-terminal devices. Furthermore, the weight modulation of IGTs can be much more linear, which can greatly reduce the overhead for weight programming compared with devices based on other materials. However, the IGT devices shown to date usually have lower compatibility with standard fabrication processes and have difficulty in the fabrication of large-scale arrays.

2. Conclusion

In conclusion, memristors based on resistive switching materials with different mechanisms have demonstrated a variety of

encouraging characteristics, such as low power consumption, high scaling potential, fast switching speed, long retention, multilevel conductance, and high-order complexity. The MVM engines, artificial neurons, artificial synapses, and sensory systems constructed using memristive devices have shown great potential in highly efficient and functional neuromorphic computing. However, depending on the specific resistive switching mechanism, technological maturity, and manufacturing cost, there are still technical challenges related to the above material systems. It is important to compensate and correct the adverse effects existing in the application of the materials through the improvement from materials, systems, and even algorithms, such as reducing the impact of conductance drift and programming noise on the accuracy of PCM based convolutional neural networks (CNNs) by improving the training algorithms.[175] On the other hand, we can also design systems and algorithms to exploit the non-ideal effects in various memristors as resources for improving computational efficiency, such as utilizing the variations and noise in memristor arrays for accelerating the convergence of Hopfield neural networks.[200] Once the physical attributes of the memristive devices are properly utilized, they can play important roles in efficient neuromorphic computing.

VII. NOVEL COMPUTING CONCEPTS

A. Embracing variability

Damien Querlioz, Louis Hutin, and Elisa Vianello

1. Status

Emerging nanoelectronic components, such as memristors and spintronic devices, offer exceptional features. However, these devices also exhibit a high degree of variability in their behavior due to their atomic-level features and reliance on sophisticated, sometimes incompletely understood, physics. This variability has made it necessary to model these devices using statistical tools, effectively treating them as random variables. Interestingly, multiple applications, particularly in machine learning and security, require random variables, which are expensive to generate using the traditional CMOS technology. Therefore, exploiting the inherent variability of these nanodevices presents a unique opportunity to develop efficient random number generators and stochastic computing models (see Fig. 16).

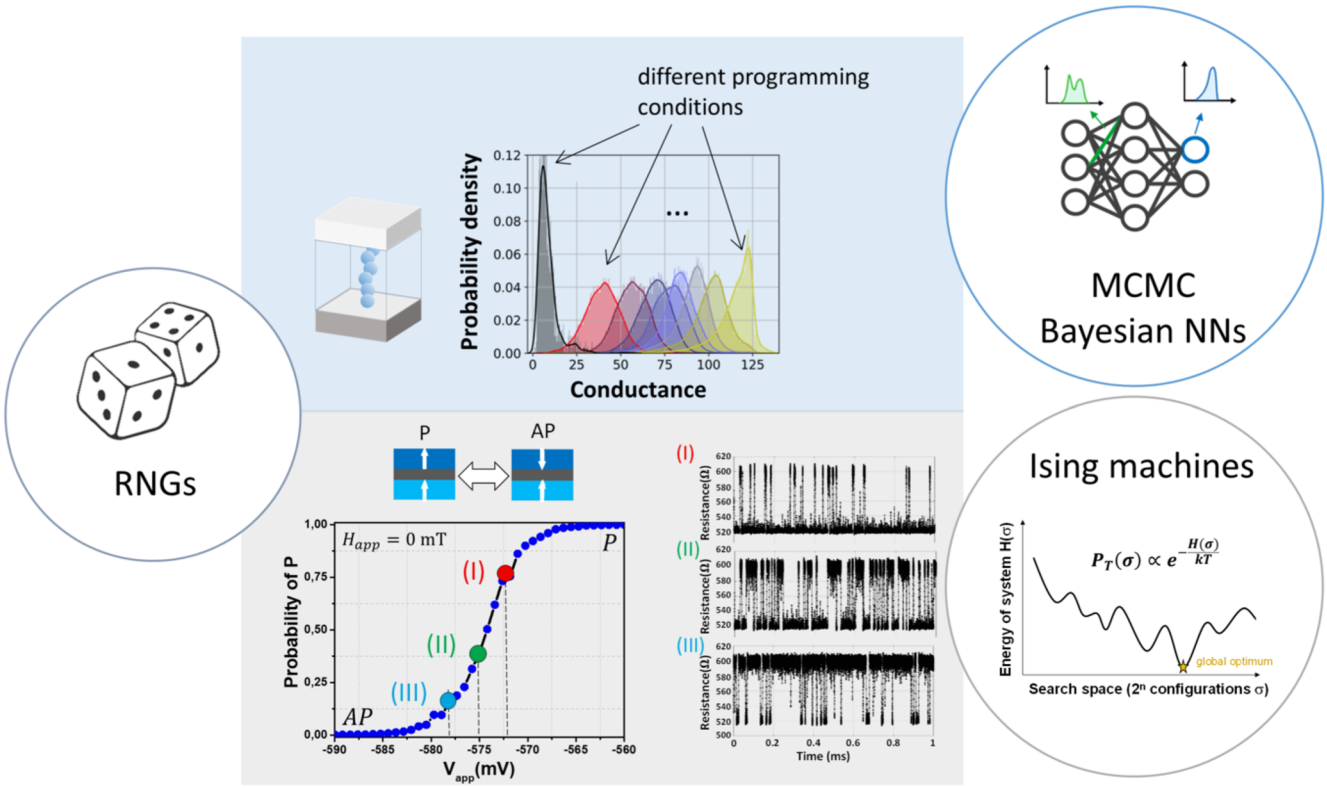

FIG. 16. Illustration of some leading approaches exploiting the variability of nanodevices for computing. Top: measurements of the programming variability of hafnium-oxide filamentary memristors. The statistical rule that this variability naturally implements can be used to perform Markov chain Monte Carlo training. Bottom: measurement of stochastic magnetic tunnel junctions, naturally implementing p-bits, used, for example, in Ising machines. Both devices can also be used for random number generation.

The first idea stems from utilizing the cycle-to-cycle read or programming variability to create random number generators (RNGs) that consume less power than traditional pseudo-random or truly random number generators. By harnessing the inherent variability in these devices, we can develop RNGs that are not only more energy-efficient but also offer improved security and robustness for various applications. Experimental realizations using filamentary memristors,[90] phase change memories,[365] and spintronic devices[366,367] have validated this concept. An interesting application is stochastic computing, an alternative approximate low area/low energy computing scheme that has been held back by the lack of compact RNGs: It requires large amounts of RNGs, which in conventional implementations dominate the area of circuits.[368,369] Stochastic computing can, therefore, strongly benefit from such stochastic nanodevices.[370,371]

The second idea involves making these RNGs adjustable, i.e., the probability for an output to be one can be controlled by an input signal, effectively creating probabilistic bits or "p-bits."[372] Such structures are analogous to stochastic binary neurons,[373,374] and they have been used in adaptive inference models designed for optimization problems, wherein a set of variables evolves through local and more-or-less random transformations toward configurations that are increasingly probable as they minimize energy.[277,372] A particularly exciting opportunity is their use within Ising machines, which have shown potential for solving highly complex tasks using reduced resources.[375] Some of the most promising p-bit implementations use spintronics, as low-energy barrier magnetic tunnel junctions provide a p-bit functionality almost intrinsically.[277,372,373] Some recent studies[376,377] have used a limited number of stochastic devices as fast high-quality randomness sources for larger FPGA-based circuits, while stressing that the projected benefits of utilizing nanodevices at a larger scale intrinsically harnessing randomness from the thermal bath remain significant and appealing. Memristors with high random telegraph noise[378] could also be used in that direction.

The third idea underscores the striking parallels between the behavior of nanodevices and the principles of Markov Chain Monte Carlo (MCMC) algorithms, a class of stochastic optimization techniques. This correspondence is particularly evident in the case of the Metropolis–Hastings MCMC algorithm. In this context, it is required to generate and store multiple random values for a single parameter, a process that can be naturally implemented using the inherent variability of nanodevices. For instance, the programming current can be employed to determine the mean value, while the imperfections of the devices provide the necessary randomness. This family of algorithms has been extensively used for sampling synaptic weight distributions in training Bayesian models, which excel at modeling uncertainty in complex situations. The most important experimental demonstration, using filamentary memristors, is presented in Ref. 197.

While MCMC algorithms serve as a compelling illustration, they are by no means the only example of stochastic computing methods that can benefit from the unique properties of nanodevices. Indeed, a diverse and rapidly evolving field of research is currently exploring the potential of other stochastic computing techniques for a wide range of applications, from neural networks to combinatorial optimization problems. As our understanding of nanodevice behavior continues to deepen, the opportunities to leverage their inherent variability in novel and transformative ways promise to fuel further innovation in stochastic computing and beyond.

2. Challenges

The primary challenge in exploiting the imperfection of nanodevices lies in the imperfect nature of the imperfections themselves. To harness the inherent variability of these devices for practical applications, it is necessary to achieve a certain level of "controlled" imperfection, which refers to maintaining the desired degree of variability without compromising the reliability and stability of the devices.

The details of memristor or superparamagnetic tunnel junction imperfections are subject to variability. This variability stems from various factors, such as manufacturing variations, environmental conditions, and the complex interplay of atomic-level features and underlying physics. In the case of superparamagnetic tunnel junctions, additional variability in time can occur due to their sensitivity to magnetic field, which is higher than in stable magnetic tunnel junctions. Magnetic shielding can be required to suppress this sensitivity.[367] Overall, modeling the imperfections of these nanodevices accurately becomes a challenging task, as capturing these variations in a consistent manner is difficult.

This challenge is further exacerbated by the fact that different types of imperfections have different impacts on the performance and usability of the devices. Therefore, identifying and understanding the specific imperfections that can be harnessed for the development of efficient RNGs and stochastic computing models is of utmost importance.

3. Potential solutions

To address the challenges associated with exploiting the imperfections in nanodevices, it is essential to acknowledge that not all imperfections are equally exploitable. For instance, cycle-to-cycle (C2C) variability and device-to-device (D2D) variability differ in terms of their usability and intrinsic nature. C2C variability arises from the inherent variability in the behavior of a single device across different operational cycles, making it more directly applicable and intrinsic to the device. On the other hand, D2D variability occurs due to variations in the performance of different devices, which may be influenced by manufacturing inconsistencies or other external factors.

C2C variability in nanodevices can manifest in two distinct forms: the variability resulting from two consecutive read operations on a device and the variability arising from two programming operations on the device. Distinguishing between these two effects can be challenging, particularly because both types of variability are highly dependent on the read and programming conditions of the devices. These two forms of C2C variability offer unique advantages from an algorithmic perspective, but they are utilized differently in various applications. For instance, read variability is especially well-suited for RNGs or probabilistic bits,[277,367] while programming variability is specifically tailored for MCMC algorithms.[197]

Even if C2C variability is, in general, more appealing than D2D variability from an algorithmic perspective, D2D variability also has applications: It enables generation of distinct and irreproducible

signatures for each device. Physically unclonable functions (PUFs) are cryptographic primitives that derive their security from the unique and unpredictable physical characteristics of individual devices and are therefore an excellent example of harnessing D2D variability for practical applications.[379]

Understanding the specific physics underlying different nanodevices can greatly impact their suitability for exploiting imperfections. For example, in hafnium-oxide memristors, only the low-resistance state (LRS) can be readily exploited for our proposed applications, as this state exhibits a much-more controlled degree of variability compared to the high-resistance state (HRS).[197] Spintronic devices also show promising potential in this regard, as they rely on physical phenomena that inherently exhibit a degree of randomness, which can be understood and modeled.[277,367,373,374,380] For example, the switching time of a magnetic tunnel junction is directly connected to the initial angle of the free layers' magnetization, a random quantity, which can be well-modeled[381] and, to some extent, controlled. When using low-barrier MTJs as artificial spins in Ising machines, D2D variability may cause an unwanted spread in threshold values or even effective pseudo-temperature across the network. It was shown that these non-ideal and non-uniform activations could be compensated to some degree by relearning the synaptic weights following the initial mapping.[382]

Another core idea for dealing with nanodevice imperfections can be to utilize multiple devices instead of relying on a single one. By employing an ensemble of devices, we can achieve better statistical properties, resulting in improved performance and robustness of the RNGs and stochastic computing models. This approach also helps mitigate the impact of individual device variations and reduces the reliance on any single device, thereby enhancing the overall reliability and stability of the systems (examples of this strategy are seen in Refs. 383–385).

4. Concluding remarks

Embracing the inherent variability of emerging nanoelectronic components, such as memristors and spintronic devices, presents a unique opportunity to advance stochastic computing, machine learning, and security applications. By understanding and exploiting the intricate relationship between the variability of these devices and the requirements of various algorithms, we can develop efficient random number generators, p-bits, and Markov chain Monte Carlo implementations, among other stochastic computing techniques.

However, not every nanodevice imperfection can be exploited. To fully harness the potential of these nanodevices, it is essential to address the challenges associated with their imperfect nature. Achieving a "controlled" level of imperfection, understanding the impact of different types of variability, and leveraging the specific physics underlying each device are crucial steps in this endeavor. By employing multiple devices in an ensemble, we can further enhance the reliability, stability, and performance of the resulting systems. As our understanding of the nanodevice behavior and the opportunities to exploit their inherent variability continues to deepen, we can anticipate a surge in innovation in stochastic computing and beyond.

B. Spiking-based computing

Sayeed Shafayet Chowdhury and Kaushik Roy

1. Status

Spiking neural networks (SNNs) are a promising energy efficient alternative to traditional artificial neural networks (ANNs). While ANN based deep learning has achieved tremendous progress in fields such as computer vision and natural language processing, it comes at a cost of huge compute requirements. Spike-based neuromorphic computing[386] provides a potential solution to this issue using brain-inspired event-driven processing. SNNs use binary spikes for computation contrary to analog values used in ANNs. A schematic of spiking neurons with their temporal dynamics is shown in Fig. 17. The spiking neuron receives spike inputs over time, which are accumulated in the membrane potential (V_{mem}), which upon crossing a threshold (V_{th}), emits an output spike.

A key characteristic of SNNs is the notion of time. While conventional feedforward ANNs are able to map static inputs to outputs with very impressive performance, learning long-term temporal correlations using them is challenging. On the other hand, recurrent neural networks (RNNs) are more suited to process temporal information efficiently,[387] although recent developments, such as transformers, have shown that ANNs can perform well in temporal processing too,[388] albeit with a higher training and memory cost. Different variants of RNNs, such as vanilla RNNs,[389] long-short term memory networks (LSTMs),[390,391] and gated recurrent units (GRUs),[392,393] have been proposed, which differ in their degree of complexity and capability to capture temporal information. However, they all contain explicit feedback connections and memory elements to handle temporal dependencies. On the contrary, SNNs can be regarded as a simpler form of RNNs, where the recurrent dynamics of V_{mem} acts as an internalized memory.[394] Interestingly, the leak in SNNs can play the role of a lightweight gating mechanism, thereby temporally filtering out some irrelevant information.[395] In addition, SNNs may lead to a lower parameter count and an easier training overhead compared to LSTMs.[396]

The sequential nature of processing in SNNs leads to unique opportunities in terms of input representation. Traditional feedforward ANNs, such as CNNs, and the more recent vision transformers[397] process several temporal inputs by merging them into a single large representation. On the other hand, SNNs can process the sequential inputs in a streaming fashion, using the inherent

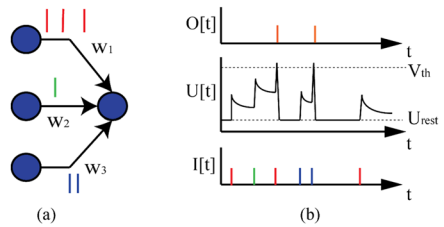

FIG. 17. A leaky-integrate-and-fire (LIF) neuron. (a) Schematic connection between three pre-neurons and a post-neuron; (b) temporal dynamics of the post-neuron. Adapted from Ref. 387.

recurrence of neuronal membrane potential. As a result, SNNs are inherently suitable to process temporal event camera[40] data. However, for analog inputs, it becomes a key challenge to efficiently encode the data into a spike train. Initial studies[398,399] use Poisson rate-coding, where the input is compared to a random number at each time step and a spike is generated if the input is higher than the random number. However, this process suffers from high inference latency. Temporal coding schemes, such as phase coding,[400] burst[401] coding, DCT-encoding,[402] time-to-first-spike (TTFS) coding,[403] and temporal coding,[404] attempt to capture the temporal correlation in the data. However, their accuracy is often lower than ANNs. More recently, direct encoding approach[405,406] has become popular, where analog values are given directly to the SNN and the first layer of the neural network layer acts as a spike generator. Such a method has provided impressive performance on complex tasks with very few time steps. Besides the analog input modalities, DVS cameras (such as DAVIS240[407]) provide discrete spikes directly as inputs, which are inherently more amenable to SNNs. Conventional ANN-based approaches tackle the event streams by accumulating them over time and subsequently processing the lumped input altogether.[408,409] However, the rich temporal cues present in the event data may not be optimally leveraged in this process. Therefore, recent studies proposed to process the event streams using SNNs,[410] which leads to a synergistic alliance between the inputs and spike-based processing. We believe that this is a promising direction to pursue as it enables harnessing the inherent temporal processing capabilities of SNNs with event data.

2. Challenges and potential solutions

A crucial bottleneck in the advancement of SNNs is the lack of suitable training methods. Sparsity of activations and the discontinuous derivative of spike functions lead to training complexity in SNNs. To counter that, initial approaches mostly used ANN–SNN conversion.[398,411] Although the conversion method provides high accuracy, it suffers from significant inference latency. Following this, surrogate gradient-based backpropagation (BP) methods have been proposed[399,412,413] to train SNNs from scratch. Note that, due to the sequential nature of inputs, SNNs are trained with backpropagation-through-time (BPTT), similar to RNNs. However, simpler neuron models and lower parameter complexity makes the optimization of SNNs simpler compared to RNNs. Although these surrogate gradient-based BPTT methods have advanced the field of SNNs by obtaining high accuracy, the training workloads are still quite intensive in addition to considerable inference latency (100 time steps). To overcome these, the authors in Ref. 414 propose to merge the conversion and BP-based training methods (termed "hybrid" training), where first, an ANN is trained to use it as initialization for subsequent surrogate gradient-based BP. More recently, advanced training approaches, such as temporal pruning,[415] custom regularizers,[416] and modified neuron models,[417] have been proposed, which enable reducing the latency of SNNs to unit time step. A complementary research direction proposes to utilize equilibrium propagation to train SNNs.[418,419] These approaches provide a promising more bio-plausible alternative to backpropagation. However, challenges remain in their large scale implementation.

Parallel to algorithmic developments, advancements in neuromorphic hardware fabrics are equally critical to unearth the true potential of SNNs. Due to the sequential nature of data processing, SNNs present unique challenges on the hardware front as current graphics processing units (GPUs) and tensor processing units (TPUs) are sub-optimal to exploit the high temporal as well as spatial sparsity.[420] Furthermore, information processing using membrane potential over multiple time steps leads to memory-intensive operations, an overhead that is non-trivial to mitigate using off-the-shelf digital accelerators. Taking such issues into account, several research directions have been pursued in recent years across the stack from devices and circuits to architectures. Event-driven neuromorphic chips, such as Neurogrid[421] and TrueNorth,[50] are notable, which are based on mixed signal analog and digital circuits, respectively. Two standout features of these neuromorphic chips are asynchronous address event representation and networks-on-chip (NOCs). Another promising direction is investigating various beyond von Neumann computing models to counter the "memory wall bottleneck." To this end, near-memory and in-memory[50,422,423] computing paradigms are being explored to improve throughput and energy efficiency. To realize these emerging computing platforms, exciting progress is being achieved in the device domain utilizing non-volatile technologies.[424] Some noteworthy approaches based on memristive technologies include resistive random-access memory (RRAM),[343] phase-change memory (PCM),[356] and spin-transfer torque magnetic random-access memory (STT-MRAM).[425] RRAMs provide analog programmable resistance but are prone to process and cycle variations and read/write endurance. Devices based on PCM can achieve comparable programming voltages and write speed to RRAMs; however, high write-current and resistance drift over time cause issues. On the other hand, compared to RRAMs and PCMs, advantages of spin devices[426] are almost unlimited endurance, lower write energy, and faster reversal. However, their ON/OFF ratio is much smaller than in PCMs and RRAMs, requiring proper algorithm/hardware co-design.[427] Note that each of these technologies has its pros and cons and there is no single winner at the moment. Floating-gate transistors[148] are another class of non-volatile devices that are being explored for synaptic storage. While their compatibility with MOS fabrication process is attractive, challenges persist regarding reduced endurance and high programming voltage.

3. Conclusion

To conclude, SNNs are a promising bio-plausible alternative to conventional deep neural networks. However, it is imperative to understand the "*why*" and "*where*" of their proper usage. While SNN algorithms have largely focused on static vision tasks[428] until now, their true potential lies in processing sequential information. To that effect, several works are exploring event-based vision for optical flow, depth estimation, egomotion, etc. We believe that immense opportunities lie ahead in further exploration of SNNs in varied avenues requiring temporal processing, such as video processing, reinforcement learning, speech, and control. In order to achieve that, there is a need of concerted synergistic efforts on algorithms as well as hardware. On the algorithmic aspect, we need to focus on investigating learning approaches that can leverage the unique data representation provided by spiking neurons. In addition, developing hardware geared toward SNN-specific algorithms is critical for the whole field to move forward. Overall, while domain-specific challenges are prevalent in spike-based computing, we believe that the next few years will be exciting as we discover the niche of

SNNs, most likely comprising temporal applications with low power requirements.

C. Analog computing for linear algebra

Zhong Sun, Piergiulio Mannocci, and Yimao Cai

1. Status

Linear algebra problems are being solved in every corner of the information world. Solving these problems by running algorithms in digital computers, however, is generally hard and resource-demanding, featuring a high computational complexity, such as $O(n^3)$, where n is the number of variables. To overcome the inadequacy of digital computers whose performance is fundamentally limited by the ultimate scaling of Moore's law and the intrinsic bottleneck of von Neumann architecture, analog computing arises as a promising solution, thanks to its efficient information encoding, massive parallelism, fast response, and the emerging resistive memory technology.[429] Analog matrix computing (AMC) is conveniently realized with a crosspoint resistive memory array, which forms a physical matrix by storing entries as crosspoint device conductances and thus can be used for linear algebra computations. There are several resistive memory device concepts that rely on distinct underlying physics, including two-terminal devices, such as resistive random-access memory (RRAM), phase change memory (PCM), magnetoresistive RAM (MRAM), and ferroelectric tunnel junction (FTJ),[430] and three-terminal devices, such as ferroelectric field-effect transistor (FeFET) and electrochemical RAM (ECRAM). They are all simply used as programmable resistive devices to implement AMC.[75,199,431–434] In this context, one of their differences lies in the conductance range that may limit the capacity of mapping matrix elements, say one bit or multiple bits. It is possible to replace one type of resistive memory device that has been demonstrated for AMC application by another one. For simplicity, we limit our discussion to RRAM that we have used frequently.

The most straightforward AMC implementation is to perform the matrix–vector multiplication (MVM) in one step. By simply applying simultaneously a set of voltages (representing an input vector) to the crosspoint columns, the currents through the crosspoint array are collected at the grounded rows, constituting the output vector, which in turn is converted and read out with transimpedance amplifiers (TIAs) [Fig. 18(a)]. By adopting the conductance compensation strategy,[435] MVM of a mixed matrix that contains negative entries can be implemented with the same number of TIAs as shown in Fig. 18(a). MVM is the backbone of many important algorithms, such as neural networks and discrete transformations. Consequently, RRAM-based AMC has widely been considered as an accelerator approach, showing more than two orders of magnitude improvements of throughput and energy efficiency.[178]

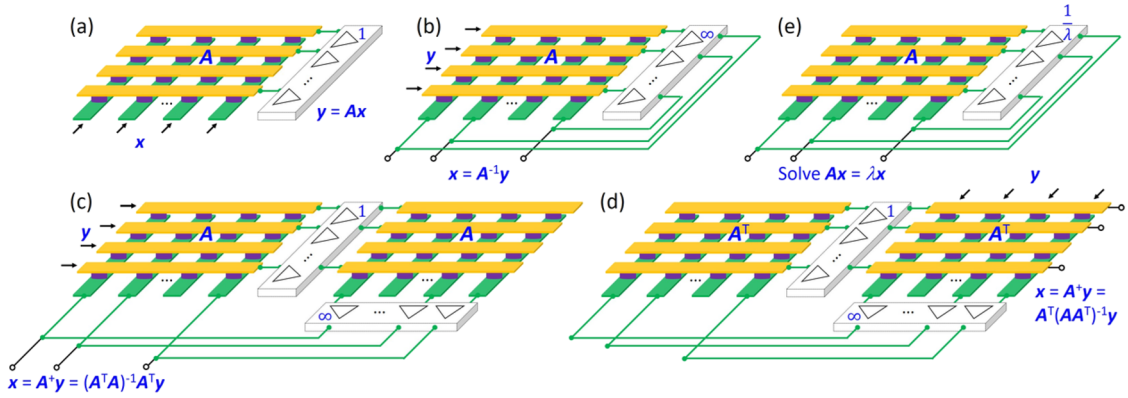

FIG. 18. AMC circuits for (a) matrix–vector multiplication, (b) matrix inversion, (c) generalized left inverse, (d) generalized right inverse, and (e) eigenvector computations.

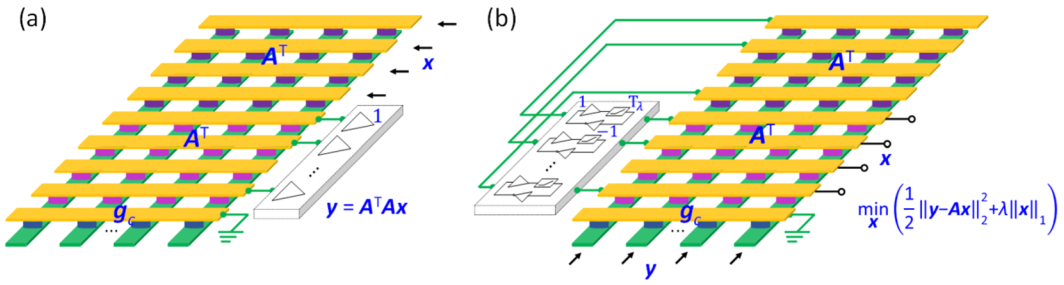

FIG. 19. AMC circuits for (a) matrix–matrix–vector multiplication and (b) solving sparse approximation problems, e.g., compressed sensing recovery.

In addition to the naive MVM application, more complicated linear algebra computations have been realized through configuring AMC circuits with feedback loops. Figure 18 shows closed-loop AMC circuits for other basic matrix operations, including matrix inversion (INV), generalized inverse, and eigenvector. The INV circuit in Fig. 18(b) is constructed based on the global feedback connections between crosspoint rows and columns through operational amplifiers (OPAs). It solves a system of linear equations when an input current vector is provided, with the output voltages of OPAs representing the solution.[436] INV is exactly the inverse problem of MVM. Both circuits utilize the same electronic components, whereas the different connection topologies define the opposite functions. The INV concept can be generalized to non-square matrices, by configuring AMC circuits with two sets of crosspoint arrays and amplifiers. Depending on the matrix shape (tall or broad), the generalized left [Fig. 18(c)] and right [Fig. 18(d)] inverse circuits have been designed, both based on the same feedback loops while showing differences in terms of input terminals and matrix storage (transpose or not).[190] INV and generalized inverse find application in many scenarios, such as machine learning, wireless communications, and scientific computing.[437–439] Unlike MVM, INV suffers from the condition number issue, where the input error would be amplified for an ill-conditioned matrix (with a large condition number), resulting in a low computing precision. Consequently, it is appropriate to use the INV circuit as an analog preconditioner, and the performance of scientific computing may be improved by more than three orders of magnitude.[437] The eigenvector circuit in Fig. 18(e) uses global feedback as well, but it is a fully self-sustained system with no external inputs, while working by positive feedback mechanism.[436] It finds application in typical scenarios, including quantum simulations, PageRank for Google search, or recommender systems.

Recently, more AMC circuits have been developed for solving more complicated matrix problems. Figure 19(a) shows a design for matrix–matrix–vector multiplication (MMVM), by mapping two matrices (or two copies of one matrix) in a RRAM array, assisted by the use of conductance compensation.[440] By connecting this MMVM circuit with other analog components, and particularly a nonlinear function module based on operational amplifiers, to form a feedback loop, the resulting circuit solves the sparse approximation problem in Fig. 19(b) in one step without discrete iterations. Notably, the nonlinear function may also be implemented by a volatile resistive switching device, thus substantially improving the compactness of the AMC circuit.[441] It has been used for compressed sensing recovery of sparse signal, natural images, and medical images, representing a highly promising solution for the backend processor to deliver real-time processing capability in the microsecond regime.

2. Challenges

Thanks to the manufacturability and compatibility of RRAM devices in modern CMOS technology processes, large-scale RRAM arrays have been fabricated for AMC implementations. In particular, RRAM macros including peripheral circuitries are usually designed to deliver MVM accelerations. By contrast, closed-loop AMC has been limited to small-scale concept demonstration, e.g., for 3 × 3 matrices. The reason behind the developmental stagnation might be ascribed to the unconventional analog circuitry. Different from MVM that consists of only local feedback, closed-loop AMC circuits contain complicated, hard-wired global feedback connections across the entire RRAM array. Since all elements are involved to provide a collective circuit response, the operation may become sensitive, risking the damage of RRAM devices under excessive electric stimulus. In addition, for large-scale circuits, the non-ideal factors of devices and circuits will jointly lead to an exaggerated deviation from the correct result. In addition, the time response of the circuit might be influenced by non-idealities, such as parasitic resistances and capacitances, which may even cause an instability issue.

Despite the lack of large-scale demonstration at this moment, it is quite promising to build closed-loop AMC circuits based on the relatively mature RRAM technology in the near term, given that lots of theoretical and simulation works have intensively examined the potential issues. In the long term, the following aspects shall be addressed to support the development of AMC circuits for linear algebra:

(1) Device/array level: At present, the largest available array size is 512 × 512.[173] In practice, only a fraction of the array is turned on for computation, due to the current/power overload and the accuracy limitation. We believe that such an array size is already sufficient to well support the advantages of AMC over other paradigms. Otherwise, we should put the stress on the analog conductance tunability of RRAM devices, which are expected to show as many distinguishable conductance levels as possible. In turn, fast and accurate programming of the RRAM conductance is essential to maximizing the AMC efficiency. The linearity and symmetry of device conductance update have been continuously emphasized for the online training of neural network weighs;[442] such a characteristic should also be favored for real-time update of matrix elements in general AMC applications. On the other hand, there should be a trade-off consideration on the RRAM conductance range, which is associated with the power consumption, alleviation, or exacerbation of the impacts of resistive and capacitive parasitics. The device variations should matter most to affect the computing precision. In particular, for solving inverse problems with the closed-loop circuits, the matrix structure related to the condition number should also be a deterministic factor.

(2) Circuit/architecture level: The scalability of AMC circuits comes at the cost of the reduced precision, where the accumulated error may eventually decline the nominal result. As a result, there is a trade-off between the desired computing accuracy and the possible array size. For solving extremely large-scale problems, especially for the inverse matrix problems, it is imperative to have algorithmic solutions to recover the result correctly. Therefore, a clever design of efficient algorithms for matrix processing would be very helpful. In particular, as RRAM-based AMC uses two physical attributes, namely conductance and voltage, the cascading of an algorithm may require transition between the two attributes, which will inevitably make the operation complicated. To this end, algorithms featuring as less such transitions as possible are precious. In addition, since AMC circuits are all based on the core RRAM array, it is valuable to have a reconfigurable architecture for performing different

matrix operations. As amplifiers play another critical role in AMC, it is highly beneficial to design efficient amplifiers that are well suited to the circuits.

(3) Software and application: During the AMC operation, multiple rows/columns of the array are simultaneously activated to carry out computation, indicating a primary requirement of such instructions in an AMC program. In addition, instructions for circuit reconfigurations should be included to enable a general AMC architecture. Therefore, the efforts on compiling and programming for AMC will be of significant importance.[443] To support the AMC concept, finding a killer application that well matches the advantages of AMC will be most convincing. The application of MVM to neural network acceleration has attracted enormous attention in past years, with a few successful silicon demonstrations toward real-world applications. For closed-loop AMC circuits, we believe that nonlinear matrix problems and related applications would be most promising, since they are more tolerant to errors as in the neural network.

3. Potential solutions

RRAM-based AMC is extremely fast in that the computation is basically a parallel reading process. To prevent the memory writing process being a bottleneck, parallel writing schemes have been developed. For instance, the weight matrix in the array can be updated by using the outer product operations, sometimes assisted by gradient decomposition methods.[444] However, these methods usually overlooked the memory nature of RRAM, that is, the analog information needs to be reliably read out. To this end, a verification circuit should be included to confine the conductance distributions. In particular, fast writing with less/no iterations will be very beneficial to saving the latency. One example is the closed-loop write scheme that uses current feedback to control the resistive switching process.[445] Because of the underlying ionic migration mechanism of RRAM devices, the endurance capability has constantly been a critical issue for both memory and AMC applications.[62] It is unlikely that this issue can be overcome solely by test method innovations; new physical mechanisms may be needed to solve it from the source. In addition, investigations on device materials and structures to fundamentally optimize analog RRAM performance and to empower the array extension are always highly desired.

The scale–accuracy trade-off of RRAM-based AMC should be elaborated to balance the efficiency and reliability. Efficient algorithms for matrix tiling and result recovery are vital for solving large-scale problems. Recently, a scalable AMC method, termed BlockAMC, has been proposed for solving large-scale linear systems. It partitions a large original matrix into smaller ones on different memory arrays and performs MVM and INV operations with the block matrices to recover the original solution.[446] To reduce the impact of non-idealities, AMC systems with both algorithmic and architectural innovations should be designed. For MVM, it is convenient to implement the bit slicing method to extend the computing precision, by using only low-precision memory devices.[447] However, as the INV circuits contain global feedback loops, it is difficult to apply this method to improve the precision of INV operations. Eventually, an analog–digital hybrid system may be required to deliver this capability. With the RRAM array as the core, the peripheral circuits and the connections can be reconfigured to perform different AMC operations. In this regard, the basic models shall include RRAM arrays, amplifiers, and reconfigurable routing. Different circuit configurations are actually hardware-embedded instructions of matrix operations.

It is of great practical convenience to have a set of fixed AMC primitives that can be adopted to realize general matrix computations, together with vector processing, parameter scaling, and variable attribute conversion. There is a trend that takes MVM and INV as two primitives to enable a reconfigurable, general-purpose AMC system, which may be used for linear regression, generalized regression, and eigen-decomposition.[448–450] Given the intrinsic noises in analog computing, we believe that nonlinear closed-loop AMC that is inherent in error tolerance is more promising toward real-world applications. Typical examples include solving some optimization problems, such as sparse coding, compressed sensing recovery, and linear/quadratic programming. These problems appear in many common scenarios, such as wireless channel estimation, magnetic resonance imaging, and signal processing.[451] In addition, front-end integration with sensors for signal processing would be encouraging, which helps save data conversions and thus reduces the accumulation of noise effects.[452] These problems typically favor relatively small RRAM arrays, thus alleviating the rigorous requirements on the device/array performance. Eventually, as the RRAM technology matures, the application to large-scale and high-precision problems will be advanced, with the help of innovative algorithms and architectures.

4. Conclusion

In the modern era, due to the strong demand for linear algebra acceleration and the rapid development of emerging resistive memory devices/architectures, analog computing has gained a renewed interest across academia and the industry. Various AMC circuits have been successfully demonstrated for fast solutions of matrix problems that constitute the basic operations for linear algebra computations. However, most of the AMC concepts still remain in the laboratory prototype stage, calling for a roadmap covering different aspects to guide system integration, optimization, and application. Next steps toward effective AMC shall include developing (1) reliable analog conductance programming methods, (2) architecture and algorithm designs for large-scale problems, (3) circuit designs for non-ideality mitigation, (4) reconfigurable systems, (5) more AMC circuits, e.g., for nonlinear matrix operations, and (6) application to near-term and long-term typical scenarios.

D. Analog content addressable memories (CAMs) for in-memory computing

Giacomo Pedretti

1. Status

Content addressable memories (CAMs) are a class of memory structures that, given an input query, return their stored location or

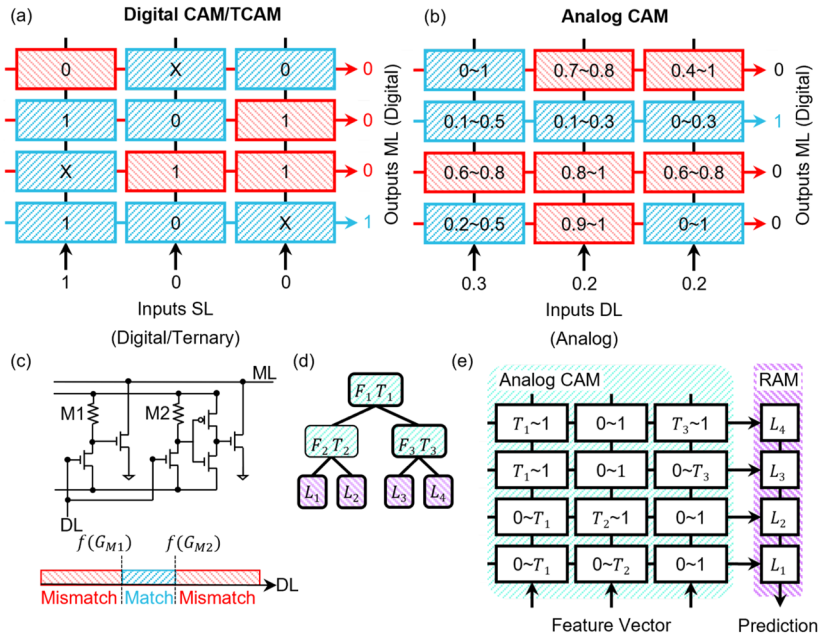

FIG. 20. Illustration of a TCAM (a) and an analog CAM (b). (c) Circuit schematic of a 6T2M analog CAM, capable of returning a match if the input on the DL is within a function of the memristor conductance (inset). (d) Example of decision tree and (e) its mapping to an in-memory computing circuit for single/few cycle inference.

address. A wildcard, "X," can be added allowing for "fuzzy" searches, resulting in a ternary CAM (TCAM).[453]

Figure 20(a) shows a TCAM schematic. An input query is applied along the columns, or search lines (SLs), and outputs are returned along the rows, or match lines (MLs). A given row is matched (returning a "1") if each value q_j of q is equal to the corresponding key-value K_{ij} stored in the row i (column j), which corresponds to performing the operation

$$\mathrm{ML}_i = \prod_j \overline{(q_j \oplus K_{ij})}.$$

CMOS/SRAM-based TCAMs are ubiquitous in networking, but they are usually bulky and power-hungry. Memristor-based TCAMs have been shown to outperform CMOS circuits, but given the relatively low speed and inefficient writing of memristors, they are currently being proposed for in-memory computing applications. In this framework, we have recently proposed an analog CAM[454] that stores *ranges* and returns a match on the MLs if the analog input is within the stored ranges or

$$\mathrm{ML}_i = \prod_j \left(T_{l,ij} \le q_j < T_{h,lj}\right),$$

where T_l and T_h are the stored upper and lower bounds thresholds, respectively. Figure 20(c) shows a conceptual representation of an analog CAM, and Fig. 20(c) shows the schematic of a 6-transistor–2-memristor (6T2M) analog CAM. Two 1T1R voltage dividers compare the analog input on their gate with the stored value in the memristor conductance, and their drain node controls either a pulldown transistor (left side) or a series of an inverter and a pull-down transistor (right side) to discharge the pre-charged ML if the input is lower than the lower bound or higher than the upper bound, respectively. For example, if the input voltage is high enough to turn on the input transistor, resulting in a voltage on the leftmost 1T1R drain lower than the pulldown threshold, the latter will not be activated. Similarly, if the same input voltage is low enough such that the input voltage of the inverter is above its threshold of the rightmost 1T1R, its pulldown transistor would not be activated. This results in a match, given that the ML would not discharge.

Other implementations of analog CAMs have been proposed based on compact ferroelectric structures, although given the absence of the inverter, a proper conditioning of the input should be performed.[455]

While the in-memory computing community in the past decade have been focused on the acceleration of deep neural networks (DNNs), tree based machine learning (ML) still outperforms DNN when processing tabular data due mainly to the presence of missing and categorical features.[456] Figure 20(d) shows a schematic representation of a decision tree (DT), which is essentially a collection of conditional branches (nodes) in a tree structure. Given the irregular structure, ensembles of DTs are not well suited for being accelerated in CPU and GPU, due to thread synchronization issues, load imbalance, and uncoalesced memory accces.[457] Recently, we have shown that DT ensembles can be mapped to analog CAMs, providing a one-cycle and conversion-less inference, given that the output is already digital and ready to be post-processed.[458] Figure 20(e) shows the DT of Fig. 20(d) mapped into an analog CAM array, where each root-to-leaf path is encoded in a row. The digital

ML outputs are connected to a conventional RAM storing the leaf values or prediction. All branches are executed in parallel, providing size-independent inference latency.

2. Challenges

Most of the challenges in building reliable analog CAMs are shared with crossbar arrays of memristors, although some circuit and device requirements are different. Similarly, a wide range of programmable levels in the memristor conductance is desirable, but different from crossbar arrays, slicing techniques for increasing precision[459] cannot be performed. Range-based analog CAMs have a unary encoding; thus, connecting multiple adjacent analog CAMs on a given row doubles the number of levels as opposed to bit-slicing in crossbar arrays, which doubles the number of bits, leading to an exponential overhead. Linear approaches for an improved analog CAM precision have recently been shown,[460] although with a significant circuit complexity overhead. Moreover, given the voltage divider operation performed directly on the memristor, its conductance needs to be stable at higher read voltages, e.g., 1 V, to be able to efficiently control the pulldown transistor.

While a small 2 × 2 array was realized in the 180 nm technology node and large array simulations at the 16 nm technology node were performed,[454] experiments on large arrays, i.e., 128 × 16, have yet to be realized. There are several concerns to be solved at the array level, for example, the impact of memristor variation and noise on the search accuracy and the impact of parasitic capacitances and resistances on a reliable operation. Moreover, due to the nonlinear input/output relationship, even defining the appropriate patterns for the program is challenging and ad hoc circuit-aware programming routines should be performed.

Finally, a full and general architecture design has yet to be performed, to feed and pull data with enough bandwidth to take advantage of the single-cycle memory lookup operation, eventually including multiple independent cores and performing operations on a large amount of data in parallel. The architecture should be programmable, not limited to tree-based ML inference, allowing multiple tasks such as associative searches,[455] and resilient to errors due to memristor variations, for example, including error correction codes routines.

3. Potential solutions

First, large arrays of CMOS-integrated analog CAMs with memristor devices should be designed and fabricated. Techniques for increasing the memristor conductance stability at relatively high reading voltages have been presented and should be used to develop an analog CAM-specific memristor stack. For example, a larger oxidation layer and/or a controlled deposition to limit the oxygen defect in it can be used to increase the set voltage of memristor devices, which could result in better stability.

Analog CAM circuits to support a higher number of bits can be designed by studying the appropriate logic functionalities in the case of a separate comparison of the least significant and most significant bits, for example by mapping different logic functions between adjacent cells.[461] In principle, it is possible to implement any kind of logic operation between adjacent CAM cells by opportunely connecting the pulldown transistors. Custom program and verify algorithms, efficiently including process variation and noise, are being developed,[462] in order to design appropriate target patterns to program in the memristor conductance while performing nonlinear operations.

Finally, a similar architecture to the crossbar array accelerators with a custom instruction set architecture (ISA) compiled from popular tools, such as sk-learn, can be implemented.[463] The analog CAM operation should be abstracted enough in order to efficiently map different workloads in a hierarchical way to the programmable accelerator, with multiple cores handling a different part of the problems, and a global network on chip accumulating the results to finalize the computation. In order to make each core operation reliable, an error detection scheme already designed for TCAMs can be adapted to the new analog CAM to efficiently re-program a given device once the state has been drifted out of the desired one.[464]

4. Conclusion

While a lot of efforts have been spent in the past years by academia and industry research for developing in memory computing structures based on linear operators, such as crosspoint arrays, recently, a renewed interest in memristor-based CAMs has arisen. Different circuits for memristive TCAMs and analog CAMs have been proposed, targeting multiple applications, but a circuit-level integration of CMOS circuitry with a BEOL-integrated memristor device along with a system-level analysis of performance at scale has yet to be shown. While engineering such required milestones, researchers should also focus on exploring new applications exploiting open-sourced circuit models, given the novelty of the idea; it is, in fact, likely that we are just scratching the surface of the potentiality of such computing primitives, given the inherited nonlinear operation performed with unprecedented speed and energy efficiency.

E. Optimization solvers

John Paul Strachan and Dmitri Strukov

1. Status

One of the best examples of a high-risk–high-reward area for novel computing is the area of optimization solvers. Here, the ambition is to offer *some* form of speedup or reduced resource requirements (memory, energy, etc.) in solving computationally expensive combinatorial optimization problems. In such problems, the task is to minimize a given cost (or energy) function by choosing the best configuration within a large dimensional space. Well-known examples include the traveling salesmen problem (finding a minimal route uniquely visiting various cities), graph coloring (coloring nodes such that connected nodes have different colors, while using the fewest colors), or training the weights in an artificial neural network (ANN). While some problem classes and instances can be solved approximately and quickly using greedy algorithms, many remain intractable and no known algorithm exists that can solve them all with polynomial resources and in polynomial time. Even a modest speedup could offer immediate benefits in practical applications across planning, wireless communication, bioinformatics, routing, finance, and many more.

The above challenge has inspired many research communities to develop both new algorithms and new physical hardware approaches, often in conjunction. A prominent example of this was the recognition that many optimization problems can be mapped to problems in physics,[465] such as finding the ground state of an Ising system of coupled up-or-down spins. This has driven interest in the so-called Ising solvers, which can have many physical realizations, including optical parametric oscillators, coupled electrical oscillators, magnetic tunnel junctions, and analog memristors.[375] From the early neural-inspired communities, John Hopfield's nonlinear networks of analog neurons turned out to also solve planted optimization problems.[466] Quantum or quantum-inspired approaches have also been pursued, leveraging the quantum adiabatic theorem, where a ground state (optimal) solution is found by evolving slowly from an initial, simple Hamiltonian toward a final Hamiltonian that encodes the desired problem.[467] Dynamical solvers are also being pursued,[468,469] where the large configuration space is explored rapidly via coupled nonlinear dynamical equations, with optimal configurations acting as attractor states.

It is important to realize that the variety of physics and quantum-inspired approaches discussed above offer primarily an algorithmic paradigm for solving optimization problems. The underlying materials, physics, and computational elements remain quite flexible. For example, traditional CPUs, GPUs, and FPGAs can be used to simulate nonlinear dynamical equations. Or electronic CMOS-based ring oscillators can be built and resistively or capacitively coupled in order to simulate the magnetically interacting spins in an Ising model. Thus, there are many possible combinations of algorithmic or computational models and the physical substrate where computations are performed. It is easy to mistake one for the other. Yet, despite such variety in possible physical realizations, some common underlying challenges emerge, which must be handled in order to successfully engineer an efficient and broadly useful optimization solver. The remaining portion of this document lists these challenges and opportunities.

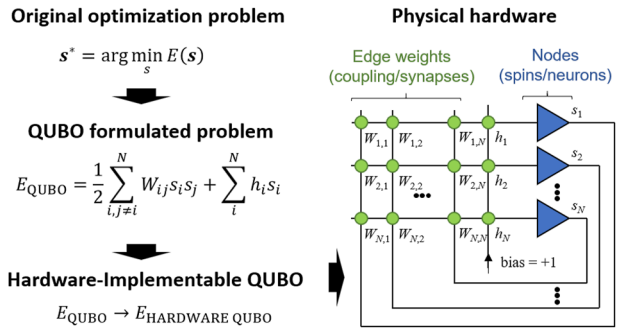

FIG. 21. Typical steps for solving combinatorial optimization problems with Ising machine. The original problem is formulated as a quadratic unconstrained optimization (QUBO) problem. The QUBO problem may need to be modified, e.g., sparsified or decomposed to smaller subproblems, before it can be mapped on the targeted Ising machine. Note that the figure shows a fully connected Ising machine and the simplest baseline QUBO formulation. Efficient implementation of different annealing techniques and approaches for solving constrained optimization problems relies on the ability to adjust coupling weights during runtime.

2. Challenges

a. Challenges of mapping to hardware. Mapping optimization problems to physical systems, such as a hardware Ising model (illustrated in Fig. 21), is an attractive approach, but it also highlights many key challenges. An Ising model is based on pairwise couplings of binary spins, and therefore, up to quadratic terms appear in an energy function with binary variables, and there is no explicit mechanism to enforce additional constraints (such as summing to certain integer values). This is also known as a Quadratic Unconstrained Binary Optimization (QUBO) type, having an energy function in the case of an Ising model,

$$E_{\mathrm{QUBO}} = \frac{1}{2}\sum_{i,j\neq i} W_{ij}s_is_j + \sum_i h_is_i.$$

Yet, many optimization problems involve higher than quadratic interactions. For example, a k-Satisfiability problem (k-SAT) involves terms of order k. Such higher-order interactions must be mapped down to quadratic terms, through the introduction of additional auxiliary variables. The result is a new total number of variables that is polynomially larger than the original number of variables, leading to, in the worst case, an exponential penalty in terms of the configuration space that needs to be searched. This has motivated explorations of algorithms with higher-order interactions[470] and their physical realizations.

In addition, QUBO formulations are sometimes inefficient, even for natively quadratic problems. For example, the typical QUBO approach for a K-city traveling salesman problem is to encode each route with K one-hot-encoded K-bit vectors representing the visitation order of each city.[465] The result is a quadratic scaling for QUBO variables, i.e., $N = K^2$, with the number of cities. This can lead to a worst-case exponential penalty in the configuration space to search as a function of the number of variables.

b. Scaling challenges. Efficient mapping ("embedding") of an optimization problem to the underlying Ising hardware may require further modification of the QUBO formulation. Indeed, the naive implementation of N-variable QUBO problems requires N^2 coupling weights to allow programmable coupling with unique weights between any pair of neurons. N can be more than a million for many practical sparsely coupled optimization problems [such as the already mentioned TSP and SAT problems], which is clearly unacceptable for most Ising model implementations. To address this challenge, a general approach is to decompose the original QUBO problem into smaller subproblems that can be implemented in the hardware. For example, one can partition a large neuron connectivity graph into smaller subgraphs, with minimized number of edges between subgraphs. The corresponding subgraph subproblems are then solved independently, e.g., by fixing values of variables not participating in the currently solved subgraph. The downside of such an approach can be lower solution quality or longer time, even if all subproblems are individually solved optimally. The scaling problem is further exacerbated for hardware approaches with limited coupling capabilities. For example, if only limited neighbor connectivity is possible—such as in the so-called chimera graph topology implemented in the DWave interconnected superconducting bits—then the embedding comes with an even higher overhead. In this case,

the original QUBO problem is first sparsified to meet the connectivity limitations by adding redundant variables and then partitioned into subproblems. The result can be an exponential slowdown[471] in solution convergence.

c. Precision challenges. Many emerging device technologies do not allow for realizing accurate weight couplings and performing precise computation of node updates. On the other hand, the weight precision should be sufficient to encode the weight dynamic range. More generally, the non-idealities in the physical couplings need to be low enough not to distort the global energy minima. Detailed studies of non-idealities in, for example, Ising machine operations are so far very limited; however, an initial insight is provided by similar studies for training artificial neural networks, a type of optimization task. Training to the highest accuracy requires at least 4-bit weight and dot-product computation precision for many deep learning models.[472] The precision requirements are higher for very compact models, such as MobileNet, and lower for larger, redundant models, such as VGG. This is likely inversely correlated with the number of global minima and their basin volumes. Therefore, higher precision requirements are expected for solving harder combinatorial optimization problems and/or when better solution quality is sought.

3. Potential solutions

Advances are needed across algorithmic, device, and architectural levels. On the algorithmic front, it was already stressed above that improved mappings are needed to ensure efficient conversion from the original optimization problem to QUBO, as well as efficient embedding.

On the device front, we need to identify the best device and material implementations to overcome challenges in connectivity and interactions. Are there any material and device concepts that would allow implementation of the higher order in the hardware? The scaling issues highlight the importance of efficiently implementing coupling weights. In this respect, emerging analog memory devices and in-memory computing with dot-product circuits are especially attractive. Optical computing approaches are attractive because of the high fan-in interconnects that can be attained. Quantum computing is attractive because of quantum annealing, which offers a tunneling mechanism through energy barriers. Many tricks on the device and circuit levels can be borrowed from the work on neuromorphic inference to improve tolerance to non-idealities and increase precision.

On the architectural side, we need more flexible designs to handle broadly varying levels of difficulty and size. Many problems are locally dense but globally sparse. For example, hard k-SAT problems are sparse, with sparseness increasing with model size, and the architecture should take advantage of such sparseness by mapping highly interconnected neuron subgraphs onto fully connected smaller size Ising machines and interconnecting Ising machines with routing networks.

A key feature for further advances on every front will be proper benchmarking. The design space is very broad and is further complicated by the wide range of applications with different characteristics (hardness, coupling density, and problem size). Figure 22 shows the key metrics—solution quality found as a function of hardware time to solution at a fixed power consumption and hardware resource budgets, highlighting the important trade-offs to identify promising solutions and guide further design. Such a figure can be drawn for a collection of benchmarks or an individual benchmark and, in the general case, would change with the scale of the problems. Note that the fixed resource budget in Fig. 22 is essential for a fair comparison of different designs because time to solution can be improved with parallelism, i.e., spending more energy and/or relying on higher complexity hardware. (Alternatively, Fig. 22 can be extended to show time, energy, and hardware complexity to the solution quality. Such characterization would be more complete, although less insightful when sparse data are available.)

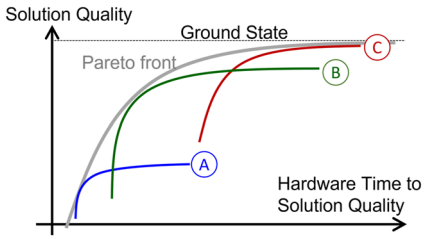

FIG. 22. Benchmarking and design space for optimization hardware (shown schematically). We expect solver implementations of type (A) that are suitable for quickly finding lower-quality solutions but incapable of reaching higher solution qualities, such as fast but approximate hardware heavily relying on parallel decomposition algorithms and/or imprecise circuits. At the other end of the spectrum are brute-force (complete) approaches (C), e.g., running an algorithm of checking all possible solutions on the conventional high-precision computer, that are slow but guaranteed to find the global optimum. Other approaches (B) may perform better for medium solution quality and time cost. We expect a Pareto front formed by different hardware approaches highlighting that linear improvements in solution comes at exponentially longer hardware times. The figure is motivated by similar dependencies at the algorithmic level (the so-called run-length distribution figures) and similar hardware trade-offs in neuromorphic inference and training.[473]

For example, we anticipate that low precision and/or restricted connectivity systems will be much faster and energy-efficient when worse precision is required. Such systems may never reach the highest quality solutions, e.g., because inherent error modifies the energy landscape or the subspace. On the other hand, a brute force algorithm, e.g., performed on a conventional computer, would guarantee finding optimal solutions. The first emerging optimization solvers (such as Ising machines) can occupy intermediate locations in these trade-off curves, tuned to the desired priority metrics.

4. Conclusions

There is great potential to harness physics- and brain-inspired approaches to improve today's computing systems for solving optimization problems. Yet, we see challenges that must be overcome at many material, circuit, system, and algorithmic levels in order to realize this potential. A variety of approaches are under exploration that use optical, quantum, magnetic, or electronic components, sometimes in combinations. However, the core requirements and issues are similar across all of them, as outlined here. In the end, it will be critical to engage other non-hardware communities to become enthusiastic users and to help in tool development for these emerging hardware systems. It will

only be through continuous feedback between the users, developers, materials scientists, and engineers that a reliable and flexible optimization solver can be developed for large-scale practical problems.

VIII. TECHNOLOGICAL MATURITY

A. Current status and next steps

Manuel Le Gallo and Stefano Ambrogio

1. Status

The various neuromorphic computing technologies and concepts covered in this roadmap have shown promising results through various prototype chip demonstrations realized by academic and industrial institutions. SRAM-based in-memory computing accelerator chips could demonstrate energy efficiencies >100 TOPS/W for 4-bit matrix–vector multiplications and provide support for all the additional CNN and LSTM inference operations with on-chip digital arithmetic and logic units (ALUs).[474] The recent analog accelerator based on NOR flash memory from Mythic, supporting 80M on-chip analog weights, demonstrated 3.3 TOPS/W system-level energy efficiency for 8-bit calculation precision and could run a pose-detection application while consuming only 3.73 W.[475] While several alternative technologies have recently been explored, such as ferroelectric and magnetic devices, 2D transistors, and memtransistors, the investigations are still at the level of single devices or small arrays, preventing a proper computation performance exploration. For this reason, in non-CMOS based implementations, only the more mature PCM and RRAM emerging technologies have been integrated into multi-core in-memory computing chips and could demonstrate various neural network inference tasks, albeit not fully end-to-end. Near software-equivalent accuracies and energy efficiencies of 10 TOPS/W or higher for matrix–vector multiplications have been reported.[56,173,476,477] Less mature technologies, such as ferroelectric and magnetic memories, have been successfully integrated into small arrays with on-chip data converters.[75] Memories based on spintronic, 2D, and atomistic materials as well as photonic processors based on resistive memories have been mainly investigated at the individual device level or integrated into small arrays without peripheral circuits performing data conversions.[297,312]

Besides chips that aim at accelerating matrix–vector multiplications for deep neural network inference tasks, platforms that can execute more novel neuromorphic computing concepts via in-memory computing have been demonstrated as well. A 64k-cell PCM chip from IBM with *in situ* learning capability using Spike-Timing-Dependent Plasticity (STPD) and leaky integrate-and-fire neurons performing a simple associative learning task was demonstrated.[478] Another 1.4M-cell PCM chip implementing a restricted Boltzmann machine with STDP learning rule could demonstrate low-power on-chip training and inference on the MNIST dataset.[170] Spiking implementations are also used to efficiently implement Residual Network (ResNET) networks on CIFAR-10.[479] Although those platforms are small prototypes and do not support end-to-end deployment of a variety of models, the key computational blocks involved in the execution of the algorithms have been successfully integrated on-chip together with the in-memory computing crossbar arrays. Nonetheless, such prototype demonstrations are still far behind the maturity of digital CMOS computing platforms, which support full deployment of a wide variety of models, often with an end-to-end software stack. CMOS neuromorphic computing chips, such as IBM TrueNorth[50] and Intel Loihi,[51] have been made available as research platforms to implement inference and training of spiking neural networks with some software support being provided. Moreover, several digital application-specific deep learning accelerators with an end-to-end software stack, such as Google Tensor Processing Unit, Amazon Inferentia, Facebook's M.2 accelerator, and IBM Artificial Intelligence Unit, have reached a mature state in terms of production and system-level integration.[480]

2. Challenges

Several challenges at the device- and system-level could hinder the development of fully end-to-end in-memory computing accelerators. The growing interest in inference tasks, where the neural network is first trained in software and then deployed on-chip to get high throughput (number of inferences per second) and low latency (time to process one input), requires chips with multiple crossbar arrays, to account for multiple layers in neural networks, and reasonably large size, to map extended network layers (e.g., first layers in convolutional networks).[476,481] This leads to several device-level requirements, such as high HRS/LRS resistance ratio, moderate endurance, high retention, and low intrinsic variability. It is also beneficial to have a LRS resistance high enough to limit the impact of the voltage drop in the lines of the crossbar array during writing and readout. Although a single device can be designed to easily meet one of those requirements, the challenge is to build multiple chips each containing multiple arrays of devices that should meet all of them. Because it is rather challenging to simultaneously achieve such specifications with emerging technologies, often trade-offs have to be made depending on the envisaged application. As an example, a lower endurance can be acceptable to achieve trade-offs for a higher HRS/LRS ratio for inference purposes. In addition, several aspects regarding the operating voltages and currents of these devices need to be considered when integrating them into crossbar arrays. Programming voltages in excess of 2 V and large programming currents (>100 μA) necessitate the use of large access transistors that can block this voltage and drive this current. Therefore, minimizing the programming voltages and currents is critical to achieve high device density in the crossbar.

On training chips, where the weights of the neural network are actively changed on-chip to get high-accuracy processing on several tasks (classification and language processing), specifications are higher. In addition to the inference requirements, data need to back-propagate through the network. Such a behavior is critical in analog cores, requiring fully symmetric peripheral circuitry during forward and backpropagation and absence of any polarity dependence of the resistive devices (since each memory element is generally biased in a different way between forward propagation and backpropagation). Clearly, device challenges are different: while inference requires high retention, low drift, and low temperature dependence, training requires high endurance and symmetric conductance update behavior, leading to different material choices.

Moreover, to achieve highly competitive system-level performance against existing CMOS-based accelerators, further improvements in device-level precision and compute density are required.[482] While this poses material challenges on resistive memories, it also influences the peripheral circuitry by limiting its available area while ensuring that it does not restrain precision. In addition, real workloads involve a variety of different operations other than matrix–vector multiplications that need to be implemented in separate digital computing units. Therefore, power-hungry analog-to-digital conversion is needed at the crossbar outputs, which limits energy efficiency. Moreover, a fast and flexible communication scheme together with highly efficient pipelining of the digital compute units and intermediate SRAM storage is primordial to ensure that they do not dominate the latency and power consumption. Finally, a user-friendly software-stack that tightly integrates such hardware with common machine learning frameworks (PyTorch/TensorFlow) is key for its widespread adoption within the community.

3. Potential solutions

The advancement in the field requires a joint development of both material aspects of the memory, algorithms used to train the neural networks, efficient network mapping techniques over multiple crossbar arrays, and careful design of the peripheral circuitry. In general, inference/training chips are required to have high-accuracy computation and high energy (operations per energy) and high area (operations per square millimeter) efficiency.

To improve the compute precision up to four or five-bit fixed-point arithmetic, it is essential to minimize the temporal conductance fluctuations (such as noise, conductance drift, and temperature dependence). To achieve this, additional materials research, such as proper incorporation of dopants, memories composed of multi-layer materials stacks,[483] and exploiting material confinement to change device properties,[357] is essential. To improve the compute density, besides scaling both the devices and the associated access transistors, high-density arrays need to be integrated at the backend of a CMOS wafer. To decrease the computational time required to convert the integrated charge/voltage after a Multiply-and-Accumulate (MAC) operation, column multiplexing should be avoided, using a per-column circuitry, shrinking the area to accommodate Analog-to-Digital Converters (ADCs). Exploration of ADC designs with a low-number of bits enables efficient integration, while still keeping a reasonably high MAC accuracy.[481] Another approach uses fully analog peripheral circuitry, which is more power efficient,[476] posing, however, more stringent limitations on computation precision and available activation functions. As an example, Rectify Linear Unit (ReLU) has been demonstrated in the analog domain, while nonlinear sigmoid or tanh generally require digital processing.

Even in the case of highly efficient crossbar arrays, the general performance could still be poor due to Amdahl's law, since fast analog cores would generate large amounts of data that the neighboring digital cores need to process.[482] In other words, highly efficient analog circuitry would require highly efficient digital blocks closely located. Therefore, to improve the hybrid analog/digital chip, spatial network mapping, analog/digital cores, spatial location, and data communication need to be co-designed, leading to a general trade-off between chip reconfigurability (ability to map any type of network) and performance (ability to efficiently process specific networks).[482]

4. Conclusion

Computing in memory has greatly improved in recent years, thanks to several on-chip demonstrations. However, challenges need to be overcome to get to product level, such as 4-to-5 bit computing precision together with performances in the range of 100/200 TOPS/W. Achieving such specifications is critical to provide a competitive advantage over existing purely digital processing cores. This will be obtained by optimizing the full computing stack: devices will require large uniformity, reasonably high resistance levels, and an overall low noise behavior. Compact peripheral circuitry will be essential to reduce the chip area, and careful spatial mapping of analog and digital dedicated cores will need to improve the performance on a variety of neural networks.

IX. EPILOGUE AND CONCLUDING REMARKS FOR THE ROADMAP

Ilia Valov, Rainer Waser, and Adnan Mehonic

After decades of reliance on transistor-based electronics, we are now delving into an era where the exploration of innovative nanoelectronic technologies based on functional materials is more crucial than ever. For instance, nanoscale memristive devices have emerged as key components for future nanoelectronics. Their straightforward stack structure, diverse functionalities, and specific benefits, such as scalability, broad temperature stability and operation range, and resilience to high-energy particles and electromagnetic interference make them indispensable for numerous applications. Furthermore, memristors and similar novel technologies now serve as foundational units for the next generation of brain-inspired computing architectures. From their introduction in the 1960s as resistive switching memories, through the relation to Chua's memristor and their use as artificial neurons and synapses at present, memristive devices have passed decades of intensive research with respect to both fundamentals and applications by academia and industry.

Examining the foundational aspects of materials science is pivotal for developing new nanoelectronic technologies. Using memristive technology as an example, the appeal of memristive devices stems from the multitude of benefits they provide, which are influenced and modulated by the materials and processes that dictate their behavior and functionalities. Here, the relation between materials, material properties, physicochemical processes, and functionalities should be particularly highlighted. A huge spectrum of materials has been used for switching films—1D (single molecules or molecular clusters), 2D (graphene, hBN, MoS_2, $MoSe_2$, etc.), and 3D, including inorganic, organic, and biomaterials. Several physical phenomena were reported to lead to memristive behavior—phase change, redox reactions and ionic transport, electronic effects, van der Waals forces, and magnetic and magneto-resistance changes, all covered in this roadmap.

All different phenomena, of course, depend strongly on the used materials. The main challenge appears to properly select a combination of appropriate materials and their dimensions (i.e., thickness and lateral scale). Apparently, a simple two-electrode cell

is composed of a switching layer or in many cases more switching layers, two electrodes, and capping film(s). The devices are exposed to extreme operating conditions, such as current densities of up to 1×10^6 A per square centimeter and electric fields of 10^8 V/m. Due to these extreme conditions and the nano-dimensions of the films, their thermodynamic and kinetic behaviors typically deviate strongly from their macroscopic counterparts going much beyond the frame of the classical knowledge in terms of chemical and mechanical stability, point defect chemistry, and transport properties. The nano-size effects are especially prominent considering the switching films. These are typically materials that, in a macroscopic sense, are insulators, which are sometimes used as high-k dielectrics, but turn into mixed solid electrolytes at thicknesses below ~50 nm. Furthermore, due to enhanced surface energy excess, the layers react chemically within the stack even if the change in the Gibbs energy for the reaction is positive. As a result, intermediate films can be formed, with similar thicknesses as the switching layers, which can either inhibit or enhance charge and mass transport through interfaces, strongly influencing all related processes. Other effects caused by impurities, absorption of moisture and/or oxygen, local changes of concentrations, and nanocavities are known to additionally complicate the control over the device's behavior and characteristics. This poses a challenge, yet also presents an opportunity to uncover unique device physics and explore alternative device functionalities (e.g., neuromorphic functionalities).

The way to keep control over the devices and to further expand the horizon of functionalities is by using a materials science-based approach where materials properties and processes are studied in very detail, and this knowledge should be applied in the design of the devices.

The "Roadmap to Neuromorphic Computing with Emerging Technologies" roadmap seeks to tackle these issues and present a contemporary perspective on the intersections between current materials science, electronic engineering, and system design. Its ultimate goal is to delve into alternative computing models, particularly brain-inspired (neuromorphic) computing. Eminent groups and experts span a wide spectrum of relevant subjects and technologies, fostering a platform for idea exchange. In addition, they highlight next-generation neuromorphic hardware, emphasizing the foundational role of functional materials and innovative device technologies. Although materials have been the main focus of the roadmap, our aim was to provide a more holistic overview and highlight a range of emerging and highly active research areas. As such, there are many details and specific material considerations not covered here that we strongly recommend the authors explore in the extensive background literature, much of which is published in special issues[13,484–486] or excellent reviews.[8,12] Likewise, in order to keep the format of the roadmap relatively compact, we have not elaborated on a number of equally valid and highly promising approaches in the context of neuromorphic technology development. Some notable examples include adaptive matters and computation based on disorder,[316,487] systems based on organic perovskites,[488,489] ionic-liquid based devices,[432,490] and molecular devices.[63,491,492] Of course, the list is not exhaustive, and other approaches, physical systems, and technologies are emerging.

Likewise, in order to keep the format of the roadmap in a relatively compact form, we have not elaborated on a number of equally valid and highly promising approaches in the context of neuromorphic technology development.

It is natural to ask what needs to happen next for neuromorphic technologies to find their way into real-world applications. Unsurprisingly, this is a complex question and heavily depends on the definition of neuromorphic and the specific application in question. In a broader sense, the integration of memory within (in-memory compute) or closer to the compute units will likely be the first step. A prime example of this is within the context of embedded systems, which are in dire need of better, more scalable non-volatile memory technology. Many memristive technologies are well-positioned to meet this demand. Among these, the most successful will likely be the technology that is the most practical (e.g., easy to embed on the same CMOS die), cost-effective, and, of course, higher performing (e.g., scalable beyond the 28 nm limit of current NOR flash, faster, more energy-efficient, and reliable with a low bit error rate). While replacing NOR flash may not be strictly neuromorphic, it will be an important step in providing better platforms to test energy-efficient neuromorphic approaches. The next step could be the development of true in-memory compute, specifically analog, with the primary goal of providing more efficient linear algebra accelerators (specifically matrix–vector multipliers). This would accelerate neural network inference engines, both in speed and in energy efficiency, and benefit other applications, such as combinatorial optimization and security. In the context of new nanoelectronic devices and analog in-memory computing, the main challenge is to achieve not only two well-separated digital states but also fully gradual analog state programmability. Variability and retention thus become significant challenges that need to be addressed. Beyond, for any analog approach to make sense, honest system-level benchmarking must be conducted, alongside an analysis of which specific target applications this approach would best serve. One of the main open questions is the translation of analog to digital signals and vice versa. This conversion process could be more costly than compute alone and might lead to overall inefficiencies compared to fully digital implementations.

Perhaps, the first two examples are not truly neuromorphic, so the next exploration, which genuinely seeks inspiration from the working principles of the human brain, involves technologies and systems aiming to implement key neurobiological features directly in hardware, whether synaptic or neuronal functionality. Other functionalities, such as dendritic operations, have also been explored. In this context, device functionalities that might typically be seen as less useful in conventional digital or analog in-memory computing may prove very helpful. For example, the volatility of states is an important feature for implementing many neuromorphic functionalities (e.g., short-term plasticity, spiking, eligibility traces, and reservoir computing).

A useful way to consider the next steps in developing neuromorphic hardware technologies is to categorize them into short-term goals (2–3 years: merging novel non-volatile memory and compute for digital systems/fabricating them on the same die at advanced processing nodes), mid-term goals (3–5 years and beyond: analog compute), and long-term goals (truly neuromorphic functionalities). While it is difficult to predict when these milestones might be reached, the first steps are already evident with MRAM, RRAM, and PCM finding their way into product offerings (e.g.,

microcontrollers with novel non-volatile memories). Recent impressive results suggest that some analog systems may be more widely utilized in the next 3–5 years, while truly neuromorphic technologies might initially be implemented in specific scenarios before achieving broader adoption (e.g., spiking systems to implement smart vision cameras). The latter two predictions still remain highly speculative.

Recommending the most promising material systems is a complex task. In general, RRAM technology might have an advantage due to its simplicity and CMOS-friendliness. However, recent developments in HfO_2-based FeRAMs represent a highly active area of research. MRAM is likely the most mature technology, with already available products, while PCM has seen significant interest recently. 2D materials are expected to integrate with all these technologies, providing further device improvements. It is important to keep in mind that the requirements might be dramatically different depending on targeted applications. For embedded systems, ease of integration and full CMOS compatibility are likely the most important factors. Conventional NVM devices need to outperform flash, while in the context of computing, higher endurance will likely be needed. In addition, the requirements for less conventional analog or neuromorphic functionalities are somewhat less defined but equally relevant and in development. One should also bear in mind the gap that exists between academic research on proof-of-concept demonstrators and industrial R&D. Industrial R&D must consider not only technical factors but also economic feasibility.[493]

Neuromorphic technologies are undoubtedly poised to be strong contenders for the future of computing, whether based on conventional digital, analog, or conceptually different computing and signal processing paradigms.

ACKNOWLEDGMENTS

We would like to express our sincere gratitude to Prof. Judith Driscoll, former Editor-in-Chief of APL Materials, for her invaluable efforts in the early stages of shaping and organizing this roadmap.

AUTHOR DECLARATIONS

Conflict of Interest

Anthony Kenyon is Founder, Director, and CSO of Intrinsic Semiconductor Technologies Ltd. Adnan Mehonic is Founder and CTO of Intrinsic Semiconductor Technologies Ltd.

Author Contributions

Adnan Mehonic: Conceptualization (lead); Project administration (lead); Supervision (lead); Writing – original draft (lead); Writing – review & editing (lead). **Daniele Ielmini**: Conceptualization (lead); Supervision (equal); Writing – original draft (equal); Writing – review & editing (equal). **Kaushik Roy**: Conceptualization (lead); Supervision (equal); Writing – original draft (equal); Writing – review & editing (equal). **Onur Mutlu**: Writing – original draft (equal); Writing – review & editing (equal). **Shahar Kvatinsky**: Writing – original draft (equal); Writing – review & editing (equal). **Teresa Serrano-Gotarredona**: Writing – original draft (equal); Writing – review & editing (equal). **Bernabe Linares-Barranco**: Writing – original draft (equal); Writing – review & editing (equal). **Sabina Spiga**: Writing – original draft (equal); Writing – review & editing (equal). **Sergey Savel'ev**: Writing – original draft (equal); Writing – review & editing (equal). **Alexander G. Balanov**: Writing – original draft (equal); Writing – review & editing (equal). **Nitin Chawla**: Writing – original draft (equal); Writing – review & editing (equal). **Giuseppe Desoli**: Writing – original draft (equal); Writing – review & editing (equal). **Gerardo Malavena**: Writing – original draft (equal); Writing – review & editing (equal). **Christian Monzio Compagnoni**: Writing – original draft (equal); Writing – review & editing (equal). **Zhongrui Wang**: Writing – original draft (equal); Writing – review & editing (equal). **J. Joshua Yang**: Writing – original draft (equal); Writing – review & editing (equal). **Ghazi Sarwat**: Writing – original draft (equal); Writing – review & editing (equal). **Syed Abu Sebastian**: Writing – original draft (equal); Writing – review & editing (equal). **Thomas Mikolajick**: Writing – original draft (equal); Writing – review & editing (equal). **Stefan Slesazeck**: Writing – original draft (equal); Writing – review & editing (equal). **Beatriz Noheda**: Writing – original draft (equal); Writing – review & editing (equal). **Bernard Dieny**: Writing – original draft (equal); Writing – review & editing (equal). **Tuo-Hung (Alex) Hou**: Writing – original draft (equal); Writing – review & editing (equal). **Akhil Varri**: Writing – original draft (equal); Writing – review & editing (equal). **Frank Brückerhoff-Plückelmann**: Writing – original draft (equal); Writing – review & editing (equal). **Wolfram Pernice**: Writing – original draft (equal); Writing – review & editing (equal). **Xixiang Zhang**: Writing – original draft (equal); Writing – review & editing (equal). **Sebastian Pazos**: Writing – original draft (equal); Writing – review & editing (equal). **Mario Lanza**: Writing – original draft (equal); Writing – review & editing (equal). **Stefan Wiefels**: Writing – original draft (equal); Writing – review & editing (equal). **Regina Dittmann**: Writing – original draft (equal); Writing – review & editing (equal). **Wing H. Ng**: Writing – original draft (equal); Writing – review & editing (equal). **Mark Buckwell**: Writing – original draft (equal); Writing – review & editing (equal). **Horatio R. J. Cox**: Writing – original draft (equal); Writing – review & editing (equal). **Daniel J. Mannion**: Writing – original draft (equal); Writing – review & editing (equal). **Anthony J. Kenyon**: Writing – original draft (equal); Writing – review & editing (equal). **Yingming Lu**: Writing – original draft (equal); Writing – review & editing (equal). **Yuchao Yang**: Writing – original draft (equal); Writing – review & editing (equal). **Damien Querlioz**: Writing – original draft (equal); Writing – review & editing (equal). **Louis Hutin**: Writing – original draft (equal); Writing – review & editing (equal). **Elisa Vianello**: Writing – original draft (equal); Writing – review & editing (equal). **Sayeed Shafayet Chowdhury**: Writing – original draft (equal); Writing – review & editing (equal). **Piergiulio Mannocci**: Writing – original draft (equal); Writing – review & editing (equal). **Yimao Cai**: Writing – original draft (equal); Writing – review & editing (equal). **Zhong Sun**: Writing – original draft (equal); Writing – review & editing (equal). **Giacomo Pedretti**: Writing – original draft (equal); Writing – review & editing (equal). **John Paul Strachan**: Writing – original draft (equal); Writing – review & editing (equal). **Dmitri Strukov**: Writing – original draft (equal); Writing – review & editing (equal). **Manuel Le Gallo**: Writing – original draft (equal); Writing – review & editing (equal). **Stefano Ambrogio**: Writing – original draft (equal); Writing – review & editing (equal). **Ilia Valov**: Writing – original draft (equal);

Writing – review & editing (equal). **Rainer Waser**: Writing – original draft (equal); Writing – review & editing (equal).

DATA AVAILABILITY

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

REFERENCES

[1] A. Mehonic and A. J. Kenyon, Nature **604**, 255 (2022).
[2] T. Gokmen and Y. Vlasov, Front. Neurosci. **10**, 333 (2016).
[3] M. J. Marinella, S. Agarwal, A. Hsia, I. Richter, R. Jacobs-Gedrim, J. Niroula, S. J. Plimpton, E. Ipek, and C. D. James, IEEE J. Emerging Sel. Top. Circuits Syst. **8**, 86 (2018).
[4] H.-Y. Chang, P. Narayanan, S. C. Lewis, N. C. P. Farinha, K. Hosokawa, C. Mackin, H. Tsai, S. Ambrogio, A. Chen, and G. W. Burr, IBM J. Res. Dev. **63**, 8:1–8:14 (2019).
[5] C. Mead, Nat. Electron. **3**, 434 (2020).
[6] A. Mehonic and A. J. Kenyon, Front. Neurosci. **10**, 57 (2016).
[7] S. Kumar, R. S. Williams, and Z. Wang, Nature **585**, 518 (2020).
[8] P. Mannocci, M. Farronato, N. Lepri, L. Cattaneo, A. Glukhov, Z. Sun, and D. Ielmini, APL Mach. Learn. **1**, 010902 (2023).
[9] D. V. Christensen, R. Dittmann, B. Linares-Barranco, A. Sebastian, M. Le Gallo, A. Redaelli, S. Slesazeck, T. Mikolajick, S. Spiga, S. Menzel, I. Valov, G. Milano, C. Ricciardi, S.-J. Liang, F. Miao, M. Lanza, T. J. Quill, S. T. Keene, A. Salleo, J. Grollier, D. Marković, A. Mizrahi, P. Yao, J. J. Yang, G. Indiveri, J. P. Strachan, S. Datta, E. Vianello, A. Valentian, J. Feldmann, X. Li, W. H. P. Pernice, H. Bhaskaran, S. Furber, E. Neftci, F. Scherr, W. Maass, S. Ramaswamy, J. Tapson, P. Panda, Y. Kim, G. Tanaka, S. Thorpe, C. Bartolozzi, T. A. Cleland, C. Posch, S. C. Liu, G. Panuccio, M. Mahmud, A. N. Mazumder, M. Hosseini, T. Mohsenin, E. Donati, S. Tolu, R. Galeazzi, M. E. Christensen, S. Holm, D. Ielmini, and N. Pryds, Neuromorphic Comput. Eng. **2**, 022501 (2022).
[10] D. Ielmini, Z. Wang, and Y. Liu, APL Mater. **9**, 050702 (2021).
[11] G. Li, L. Deng, H. Tang, G. Pan, Y. Tian, K. Roy, and W. Maass, Proc. IEEE **112**, 544 (2023).
[12] M.-K. Kim, Y. Park, I.-J. Kim, and J.-S. Lee, iScience **23**, 101846 (2020).
[13] I. K. Schuller, A. Frano, R. C. Dynes, A. Hoffmann, B. Noheda, C. Schuman, A. Sebastian, and J. Shen, Appl. Phys. Lett. **120**, 140401 (2022).
[14] F. Torres, A. C. Basaran, and I. K. Schuller, Adv. Mater. **35**, 2205098 (2023).
[15] A. W. Burks, H. H. Goldstine, and J. Neumann, *The Origins of Digital Computers* (Springer, Berlin, Heidelberg, 1982), pp. 399–413.
[16] R. R. Schaller, IEEE Spectrum **34**, 52 (1997).
[17] M. J. Flynn, Proc. IEEE **54**, 1901 (1966).
[18] R. H. Dennard, F. H. Gaensslen, H.-N. Yu, V. L. Rideout, E. Bassous, and A. R. LeBlanc, IEEE J. Solid-State Circuits **9**, 256 (1974).
[19] H. Esmaeilzadeh, E. Blem, R. S. Amant, K. Sankaralingam, and D. Burger, in *Proceedings of the 38th Annual International Symposium on Computer Architecture* (ACM, 2011).
[20] H.-S. P. Wong, H.-Y. Lee, S. Yu, Y.-S. Chen, Y. Wu, P.-S. Chen, B. Lee, F. T. Chen, and M.-J. Tsai, Proc. IEEE **100**, 1951 (2012).
[21] M. Horowitz, in *2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)* (IEEE, 2014).
[22] A. Pedram, S. Richardson, M. Horowitz, S. Galal, and S. Kvatinsky, IEEE Des. Test **34**, 39 (2017).
[23] A. Boroumand, S. Ghose, Y. Kim, R. Ausavarungnirun, E. Shiu, R. Thakur, D. Kim, A. Kuusela, A. Knies, P. Ranganathan, and O. Mutlu, ACM SIGPLAN Not. **53**, 316 (2018).
[24] A. Boroumand, S. Ghose, B. Akin, R. Narayanaswami, G. F. Oliveira, X. Ma, E. Shiu, and O. Mutlu, in *2021 30th International Conference on Parallel Architectures and Compilation Techniques (PACT)* (IEEE, 2021).
[25] S. Palacharla and R. E. Kessler, in *Proceedings of 21 International Symposium on Computer Architecture (ISCA'94)* (Los Alamitos, CA, USA, IEEE Computer Society Press, 1994), pp. 24–33.
[26] R. Bera, K. Kanellopoulos, S. Balachandran, D. Novo, A. Olgun, M. Sadrosadat, and O. Mutlu, in *2022 55th IEEE/ACM International Symposium on Microarchitecture (MICRO)* (IEEE, 2022).
[27] S. Srinath, O. Mutlu, H. Kim, and Y. N. Patt, in *2007 IEEE 13th International Symposium on High Performance Computer Architecture* (IEEE, 2007).
[28] *Supercomput. Front. Innovations* 1(1), (full issue) (2014).
[29] O. Mutlu, S. Ghose, J. Gómez-Luna, and R. Ausavarungnirun, Emerging Computing: From Devices to Systems (Springer Nature, Singapore, 2022), pp. 171–243.
[30] F. Devaux, in *2019 IEEE Hot Chips 31 Symposium (HCS)* (IEEE, 2019).
[31] Y.-C. Kwon, S. H. Lee, J. Lee, S.-H. Kwon, J. M. Ryu, J.-P. Son, O. Seongil, H.-S. Yu, H. Lee, S. Y. Kim, Y. Cho, J. G. Kim, J. Choi, H.-S. Shin, J. Kim, B. S. Phuah, H. M. Kim, M. J. Song, A. Choi, D. Kim, S. Y. Kim, E.-B. Kim, D. Wang, S. Kang, Y. Ro, S. Seo, J. H. Song, J. Youn, K. Sohn, and N. S. Kim, in *2021 IEEE International Solid-State Circuits Conference (ISSCC)* (IEEE, 2021).
[32] J. Borghetti, G. S. Snider, P. J. Kuekes, J. J. Yang, D. R. Stewart, and R. S. Williams, Nature **464**, 873 (2010).
[33] S. Kvatinsky, D. Belousov, S. Liman, G. Satat, N. Wald, E. G. Friedman, A. Kolodny, and U. C. Weiser, IEEE Trans. Circuits Syst. **61**, 895 (2014).
[34] V. Seshadri, D. Lee, T. Mullins, H. Hassan, A. Boroumand, J. Kim, M. A. Kozuch, O. Mutlu, P. B. Gibbons, and T. C. Mowry, in *Proceedings of the 50th Annual IEEE/ACM International Symposium on Microarchitecture* (ACM, 2017).
[35] A. Shafiee, A. Nag, N. Muralimanohar, R. Balasubramonian, J. P. Strachan, M. Hu, R. S. Williams, and V. Srikumar, ACM SIGARCH Comput. Archit. News **44**, 14 (2016).
[36] N. Hajinazar, G. F. Oliveira, S. Gregorio, J. D. Ferreira, N. M. Ghiasi, M. Patel, M. Alser, S. Ghose, J. Gómez-Luna, and O. Mutlu, in *Proceedings of the 26th ACM International Conference on Architectural Support for Programming Languages and Operating Systems* (ACM, 2021).
[37] M. Prezioso, F. Merrikh-Bayat, B. D. Hoskins, G. C. Adam, K. K. Likharev, and D. B. Strukov, Nature **521**, 61 (2015).
[38] C. Posch, T. Serrano-Gotarredona, B. Linares-Barranco, and T. Delbruck, Proc. IEEE **102**, 1470 (2014).
[39] T. Serrano-Gotarredona and B. Linares-Barranco, IEEE J. Solid-State Circuits **48**, 827 (2013).
[40] P. Lichtsteiner, C. Posch, and T. Delbruck, IEEE J. Solid-State Circuits **43**, 566 (2008).
[41] D. P. Moeys, C. Li, J. N. P. Martel, S. Bamford, L. Longinotti, V. Motsnyi, D. S. S. Bello, and T. Delbruck, in *2017 IEEE International Symposium on Circuits and Systems (ISCAS)* (IEEE, 2017).
[42] T. Finateu, A. Niwa, D. Matolin, K. Tsuchimoto, A. Mascheroni, E. Reynaud, P. Mostafalu, F. Brady, L. Chotard, F. LeGoff, H. Takahashi, H. Wakabayashi, Y. Oike, and C. Posch, in *2020 IEEE International Solid-State Circuits Conference (ISSCC)* (IEEE, 2020).
[43] A. van Schaik and S.-C. Liu, in *2005 IEEE International Symposium on Circuits and Systems* (IEEE, 2005).
[44] S.-W. Chiu and K.-T. Tang, Sensors **13**, 14214 (2013).
[45] S. Caviglia, M. Valle, and C. Bartolozzi, in *2014 IEEE International Symposium on Circuits and Systems (ISCAS)* (IEEE, 2014).
[46] J. A. Perez-Carrasco, B. Zhao, C. Serrano, B. Acha, T. Serrano-Gotarredona, S. Chen, and B. Linares-Barranco, IEEE Trans. Pattern Anal. Mach. Intell. **35**, 2706 (2013).
[47] E. Painkras, L. A. Plana, J. Garside, S. Temple, F. Galluppi, C. Patterson, D. R. Lester, A. D. Brown, and S. B. Furber, IEEE J. Solid-State Circuits **48**, 1943 (2013).
[48] H. Markram, Sci. Am. **306**, 50 (2012).
[49] S. Millner, A. Hartel, J. Schemmel, and K. Meier, "Towards biologically realistic multi-compartment neuron model emulation in analog VLSI," in *Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, Bruges, Belgium, 25-27 April 2012 (ESANN, 2012), https://www.esann.org/sites/default/files/proceedings/legacy/es2012-44.pdf.
[50] P. A. Merolla, J. V. Arthur, R. Alvarez-Icaza, A. S. Cassidy, J. Sawada, F. Akopyan, B. L. Jackson, N. Imam, C. Guo, Y. Nakamura, B. Brezzo, I. Vo, S. K.

Esser, R. Appuswamy, B. Taba, A. Amir, M. D. Flickner, W. P. Risk, R. Manohar, and D. S. Modha, Science **345**, 668 (2014).
[51]M. Davies, N. Srinivasa, T.-H. Lin, G. Chinya, Y. Cao, S. H. Choday, G. Dimou, P. Joshi, N. Imam, S. Jain, Y. Liao, C.-K. Lin, A. Lines, R. Liu, D. Mathaikutty, S. McCoy, A. Paul, J. Tse, G. Venkataramanan, Y.-H. Weng, A. Wild, Y. Yang, and H. Wang, IEEE Micro **38**, 82 (2018).
[52]See https://www.intel.com/content/www/us/en/newsroom/news/intel-scales-neuromorphic-research-system-100-million-neurons.html for Intel report.
[53]Y. LeCun, Y. Bengio, and G. Hinton, Nature **521**, 436 (2015).
[54]A. Tavanaei, M. Ghodrati, S. R. Kheradpisheh, T. Masquelier, and A. Maida, Neural Networks **111**, 47 (2019).
[55]L. A. Camunas-Mesa, E. Vianello, C. Reita, T. Serrano-Gotarredona, and B. Linares-Barranco, IEEE J. Emerging Sel. Top. Circuits Syst. **12**, 898 (2022).
[56]W. Wan, R. Kubendran, C. Schaefer, S. B. Eryilmaz, W. Zhang, D. Wu, S. Deiss, P. Raina, H. Qian, B. Gao, S. Joshi, H. Wu, H.-S. P. Wong, and G. Cauwenberghs, Nature **608**, 504 (2022).
[57]B. Linares-Barranco, Nat. Electron. **1**, 100 (2018).
[58]S. Furber, J. Neural Eng. **13**, 051001 (2016).
[59]H. Hendy and C. Merkel, J. Electron. Imaging **31**, 010901 (2022).
[60]A. Sebastian, M. L. Gallo, R. Khaddam-Aljameh, and E. Eleftheriou, Nat. Nanotechnol. **15**, 529 (2020).
[61]S. Brivio, S. Spiga, and D. Ielmini, Neuromorphic Comput. Eng. **2**, 042001 (2022).
[62]G. W. Burr, R. M. Shelby, A. Sebastian, S. Kim, S. Kim, S. Sidler, K. Virwani, M. Ishii, P. Narayanan, A. Fumarola, L. L. Sanches, I. Boybat, M. Le Gallo, K. Moon, J. Woo, H. Hwang, and Y. Leblebici, Adv. Phys.: X **2**, 89 (2016).
[63]V. K. Sangwan and M. C. Hersam, Nat. Nanotechnol. **15**, 517 (2020).
[64]Z. Wang, H. Wu, G. W. Burr, C. S. Hwang, K. L. Wang, Q. Xia, and J. J. Yang, Nat. Rev. Mater. **5**, 173 (2020).
[65]J. Grollier, D. Querlioz, K. Y. Camsari, K. Everschor-Sitte, S. Fukami, and M. D. Stiles, Nat. Electron. **3**, 360 (2020).
[66]G. Milano, G. Pedretti, K. Montano, S. Ricci, S. Hashemkhani, L. Boarino, D. Ielmini, and C. Ricciardi, Nat. Mater. **21**, 195 (2021).
[67]G. W. Burr, R. M. Shelby, A. Sebastian, S. Kim, S. Kim, S. Sidler, K. Virwani, M. Ishii, P. Narayanan, A. Fumarola *et al.*, Adv. Phys.: X **2**, 89 (2017).
[68]E. Covi, H. Mulaosmanovic, B. Max, S. Slesazeck, and T. Mikolajick, Neuromorphic Comput. Eng. **2**, 012002 (2022).
[69]A. A. Talin, Y. Li, D. A. Robinson, E. J. Fuller, and S. Kumar, Adv. Mater. **35**, 2204771 (2022).
[70]Y. van de Burgt, A. Melianas, S. T. Keene, G. Malliaras, and A. Salleo, Nat. Electron. **1**, 386 (2018).
[71]R. Wang, J.-Q. Yang, J.-Y. Mao, Z.-P. Wang, S. Wu, M. Zhou, T. Chen, Y. Zhou, and S.-T. Han, Adv. Intell. Syst. **2**, 2000055 (2020).
[72]K. Zhu, S. Pazos, F. Aguirre, Y. Shen, Y. Yuan, W. Zheng, O. Alharbi, M. A. Villena, B. Fang, X. Li, A. Milozzi, M. Farronato, M. Muñoz-Rojo, T. Wang, R. Li, H. Fariborzi, J. B. Roldan, G. Benstetter, X. Zhang, H. Alshareef, T. Grasser, H. Wu, D. Ielmini, and M. Lanza, Nature **618**, 57 (2023).
[73]B. J. Shastri, A. N. Tait, T. F. de Lima, W. H. P. Pernice, H. Bhaskaran, C. D. Wright, and P. R. Prucnal, Nat. Photonics **15**, 102 (2021).
[74]W. Zhou, N. Farmakidis, J. Feldmann, X. Li, J. Tan, Y. He, C. D. Wright, W. H. P. Pernice, and H. Bhaskaran, MRS Bull. **47**, 502 (2022).
[75]S. Jung, H. Lee, S. Myung, H. Kim, S. K. Yoon, S.-W. Kwon, Y. Ju, M. Kim, W. Yi, S. Han, B. Kwon, B. Seo, K. Lee, G.-H. Koh, K. Lee, Y. Song, C. Choi, D. Ham, and S. J. Kim, Nature **601**, 211 (2022).
[76]M. Payvand, F. Moro, K. Nomura, T. Dalgaty, E. Vianello, Y. Nishi, and G. Indiveri, Nat. Commun. **13**, 5793 (2022).
[77]F. Moro, E. Hardy, B. Fain, T. Dalgaty, P. Clémençon, A. De Prà, E. Esmanhotto, N. Castellani, F. Blard, F. Gardien, T. Mesquida, F. Rummens, D. Esseni, J. Casas, G. Indiveri, M. Payvand, and E. Vianello, Nat. Commun. **13**, 3506 (2022).
[78]S. Brivio, D. Conti, M. V. Nair, J. Frascaroli, E. Covi, C. Ricciardi, G. Indiveri, and S. Spiga, Nanotechnology **30**, 015102 (2018).
[79]Y. Demirag, F. Moro, T. Dalgaty, G. Navarro, C. Frenkel, G. Indiveri, E. Vianello, and M. Payvand, in *2021 IEEE International Symposium on Circuits and Systems (ISCAS)* (IEEE, 2021).
[80]L. G. Wright, T. Onodera, M. M. Stein, T. Wang, D. T. Schachter, Z. Hu, and P. L. McMahon, Nature **601**, 549 (2022).
[81]G. Tanaka, T. Yamane, J. B. Héroux, R. Nakane, N. Kanazawa, S. Takeda, H. Numata, D. Nakano, and A. Hirose, Neural Networks **115**, 100 (2019).
[82]S. Ghosh, K. Nakajima, T. Krisnanda, K. Fujii, and T. C. H. Liew, Adv. Quantum Technol. **4**, 2100053 (2021).
[83]G. Csaba and W. Porod, Appl. Phys. Rev. **7**, 011302 (2020).
[84]K. Yamazaki, V.-K. Vo-Ho, D. Bulsara, and N. Le, Brain Sci. **12**, 863 (2022).
[85]J. Zhu, T. Zhang, Y. Yang, and R. Huang, Appl. Phys. Rev. **7**, 011312 (2020).
[86]D. Ielmini and S. Ambrogio, Nanotechnology **31**, 092001 (2019).
[87]G. Zhou, Z. Wang, B. Sun, F. Zhou, L. Sun, H. Zhao, X. Hu, X. Peng, J. Yan, H. Wang, W. Wang, J. Li, B. Yan, D. Kuang, Y. Wang, L. Wang, and S. Duan, Adv. Electron. Mater. **8**, 2101127 (2022).
[88]M. D. Pickett, G. Medeiros-Ribeiro, and R. S. Williams, Nat. Mater. **12**, 114 (2012).
[89]Z. Wang, S. Joshi, S. E. Savel'ev, H. Jiang, R. Midya, P. Lin, M. Hu, N. Ge, J. P. Strachan, Z. Li, Q. Wu, M. Barnell, G.-L. Li, H. L. Xin, R. S. Williams, Q. Xia, and J. J. Yang, Nat. Mater. **16**, 101 (2016).
[90]H. Jiang, D. Belkin, S. E. Savel'ev, S. Lin, Z. Wang, Y. Li, S. Joshi, R. Midya, C. Li, M. Rao, M. Barnell, Q. Wu, J. J. Yang, and Q. Xia, Nat. Commun. **8**, 882 (2017).
[91]Y. Ushakov, A. Balanov, and S. Savel'ev, Chaos, Solitons Fractals **145**, 110803 (2021).
[92]Z. Wang, S. Joshi, S. Savel'ev, W. Song, R. Midya, Y. Li, M. Rao, P. Yan, S. Asapu, Y. Zhuo, H. Jiang, P. Lin, C. Li, J. H. Yoon, N. K. Upadhyay, J. Zhang, M. Hu, J. P. Strachan, M. Barnell, Q. Wu, H. Wu, R. S. Williams, Q. Xia, and J. J. Yang, Nat. Electron. **1**, 137 (2018).
[93]A. Mehonic, A. Sebastian, B. Rajendran, O. Simeone, E. Vasilaki, and A. J. Kenyon, Adv. Intell. Syst. **2**, 2000085 (2020).
[94]D. Kim, B. Jeon, Y. Lee, D. Kim, Y. Cho, and S. Kim, Appl. Phys. Lett. **121**, 010501 (2022).
[95]Y. Ushakov, A. Akther, P. Borisov, D. Pattnaik, S. Savel'ev, and A. G. Balanov, Chaos, Solitons Fractals **149**, 110997 (2021).
[96]A. M. Wojtusiak, A. G. Balanov, and S. E. Savel'ev, Chaos, Solitons Fractals **142**, 110383 (2021).
[97]D. Sussillo, Curr. Opin. Neurobiol. **25**, 156 (2014).
[98]S. Gepshtein, A. S. Pawar, S. Kwon, S. Savel'ev, and T. D. Albright, Sci. Adv. **8**, eabl5865 (2022).
[99]J. D. Monaco, K. Rajan, and G. M. Hwang, arXiv:2105.07284 (2021).
[100]S. Kumar, X. Wang, J. P. Strachan, Y. Yang, and W. D. Lu, Nat. Rev. Mater. **7**, 575 (2022).
[101]J. Feldmann, N. Youngblood, M. Karpov, H. Gehring, X. Li, M. Stappers, M. Le Gallo, X. Fu, A. Lukashchuk, A. S. Raja, J. Liu, C. D. Wright, A. Sebastian, T. J. Kippenberg, W. H. P. Pernice, and H. Bhaskaran, Nature **589**, 52 (2021).
[102]X. Wang, Y. Chen, H. Xi, H. Li, and D. Dimitrov, IEEE Electron Device Lett. **30**, 294 (2009).
[103]M. Mansueto, A. Chavent, S. Auffret, I. Joumard, L. Vila, R. C. Sousa, L. D. Buda-Prejbeanu, I. L. Prejbeanu, and B. Dieny, Nanoscale **13**, 11488 (2021).
[104]A. Hirohata, K. Yamada, Y. Nakatani, I.-L. Prejbeanu, B. Diény, P. Pirro, and B. Hillebrands, J. Magn. Magn. Mater. **509**, 166711 (2020).
[105]A. H. Jaafar, R. J. Gray, E. Verrelli, M. O'Neill, S. M. Kelly, and N. T. Kemp, Nanoscale **9**, 17091 (2017).
[106]L. E. Srouji, A. Krishnan, R. Ravichandran, Y. Lee, M. On, X. Xiao, and S. J. B. Yoo, APL Photonics **7**, 051101 (2022).
[107]J. C. Gartside, K. D. Stenning, A. Vanstone, H. H. Holder, D. M. Arroo, T. Dion, F. Caravelli, H. Kurebayashi, and W. R. Branford, Nat. Nanotechnol. **17**, 460 (2022).
[108]D. D. Yaremkevich, A. V. Scherbakov, L. De Clerk, S. M. Kukhtaruk, A. Nadzeyka, R. Campion, A. W. Rushforth, S. Savel'ev, A. G. Balanov, and M. Bayer, Nat. Commun. **14**, 8296 (2023).
[109]K. Vandoorne, P. Mechet, T. Van Vaerenbergh, M. Fiers, G. Morthier, D. Verstraeten, B. Schrauwen, J. Dambre, and P. Bienstman, Nat. Commun. **5**, 3541 (2014).
[110]D. J. Gauthier, E. Bollt, A. Griffith, and W. A. S. Barbosa, Nat. Commun. **12**, 5564 (2021).

[111]S.-i. Yi, J. D. Kendall, R. S. Williams, and S. Kumar, Nat. Electron. **6**, 45 (2022).
[112]A. Acín, I. Bloch, H. Buhrman, T. Calarco, C. Eichler, J. Eisert, D. Esteve, N. Gisin, S. J. Glaser, F. Jelezko, S. Kuhr, M. Lewenstein, M. F. Riedel, P. O. Schmidt, R. Thew, A. Wallraff, I. Walmsley, and F. K. Wilhelm, New J. Phys. **20**, 080201 (2018).
[113]V. Dunjko and H. J. Briegel, Rep. Prog. Phys. **81**, 074001 (2018).
[114]P. Mujal, R. Martínez-Peña, J. Nokkala, J. García-Beni, G. L. Giorgi, M. C. Soriano, and R. Zambrini, Adv. Quantum Technol. **4**, 2100027 (2021).
[115]P. Pfeiffer, I. L. Egusquiza, M. Di Ventra, M. Sanz, and E. Solano, Sci. Rep. **6**, 29507 (2016).
[116]J. Salmilehto, F. Deppe, M. Di Ventra, M. Sanz, and E. Solano, Sci. Rep. **7**, 42044 (2017).
[117]M. Spagnolo, J. Morris, S. Piacentini, M. Antesberger, F. Massa, A. Crespi, F. Ceccarelli, R. Osellame, and P. Walther, Nat. Photonics **16**, 318 (2022).
[118]K. Friston, Nat. Rev. Neurosci. **11**, 127 (2010).
[119]N. B. Janson and C. J. Marsden, Sci. Rep. **7**, 17007 (2017).
[120]H.-T. Zhang, T. J. Park, A. N. M. N. Islam, D. S. J. Tran, S. Manna, Q. Wang, S. Mondal, H. Yu, S. Banik, S. Cheng, H. Zhou, S. Gamage, S. Mahapatra, Y. Zhu, Y. Abate, N. Jiang, S. K. R. S. Sankaranarayanan, A. Sengupta, C. Teuscher, and S. Ramanathan, Science **375**, 533 (2022).
[121]M. J. Colbrook, V. Antun, and A. C. Hansen, Proc. Natl. Acad. Sci. U. S. A. **119**, e2107151119 (2022).
[122]T. Song, W. Rim, S. Park, Y. Kim, J. Jung, G. Yang, S. Baek, J. Choi, B. Kwon, Y. Lee, S. Kim, G. Kim, H.-S. Won, J.-H. Ku, S. S. Paak, E. S. Jung, S. S. Park, and K. Kim, in *2016 IEEE International Solid-State Circuits Conference (ISSCC)* (IEEE, 2016).
[123]T. Song, J. Jung, W. Rim, H. Kim, Y. Kim, C. Park, J. Do, S. Park, S. Cho, H. Jung, B. Kwon, H.-S. Choi, J. S. Choi, and J. S. Yoon, in *2018 IEEE International Solid-State Circuits Conference (ISSCC)* (IEEE, 2018).
[124]G. Yeap, X. Chen, B. R. Yang, C. P. Lin, F. C. Yang, Y. K. Leung, D. W. Lin, C. P. Chen, K. F. Yu, D. H. Chen, C. Y. Chang, S. S. Lin, H. K. Chen, P. Hung, C. S. Hou, Y. K. Cheng, J. Chang, L. Yuan, C. K. Lin, C. C. Chen, Y. C. Yeo, M. H. Tsai, Y. M. Chen, H. T. Lin, C. O. Chui, K. B. Huang, W. Chang, H. J. Lin, K. W. Chen, R. Chen, S. H. Sun, Q. Fu, H. T. Yang, H. L. Shang, H. T. Chiang, C. C. Yeh, T. L. Lee, C. H. Wang, S. L. Shue, C. W. Wu, R. Lu, W. R. Lin, J. Wu, F. Lai, P. W. Wang, Y. H. Wu, B. Z. Tien, Y. C. Huang, L. C. Lu, J. He, Y. Ku, J. Lin, M. Cao, T. S. Chang, S. M. Jang, H. C. Lin, Y. C. Peng, J. Y. Sheu, and M. Wang, in *2019 IEEE International Electron Devices Meeting (IEDM)* (IEEE, 2019).
[125]Y. Kim, C. Ong, A. M. Pillai, H. Jagadeesh, G. Baek, I. Rajwani, Z. Guo, and E. Karl, IEEE J. Solid-State Circuits **58**, 1087 (2022).
[126]F. Dong, X. Si, and M.-F. Chang, in *2022 IEEE 16th International Conference on Solid-State Integrated Circuit Technology (ICSICT)* (IEEE, 2022).
[127]X. Si, M.-F. Chang, W.-S. Khwa, J.-J. Chen, J.-F. Li, X. Sun, R. Liu, S. Yu, H. Yamauchi, and Q. Li, IEEE Trans. Circuits Syst. **66**, 4172 (2019).
[128]Y. Wang, S. Zhang, Y. Li, J. Chen, W. Zhao, and Y. Ha, IEEE Trans. Very Large Scale Integr. Syst. **31**, 684 (2023).
[129]Q. Dong, S. Jeloka, M. Saligane, Y. Kim, M. Kawaminami, A. Harada, S. Miyoshi, M. Yasuda, D. Blaauw, and D. Sylvester, IEEE J. Solid-State Circuits **53**, 1006 (2018).
[130]B. Pan, G. Wang, H. Zhang, W. Kang, and W. Zhao, IEEE Trans. Circuits Syst. **69**, 3044 (2022).
[131]J. Zhang, Z. Wang, and N. Verma, IEEE J. Solid-State Circuits **52**, 915 (2017).
[132]A. S. Rekhi, B. Zimmer, N. Nedovic, N. Liu, R. Venkatesan, M. Wang, B. Khailany, W. J. Dally, and C. T. Gray, in *Proceedings of the 56th Annual Design Automation Conference 2019* (ACM, 2019).
[133]S. K. Gonugondla, C. Sakr, H. Dbouk, and N. R. Shanbhag, in *Proceedings of the 39th International Conference on Computer-Aided Design* (ACM, 2020).
[134]B. Murmann, IEEE Trans. Very Large Scale Integr. Syst. **29**, 3 (2021).
[135]S. Spetalnick and A. Raychowdhury, IEEE Trans. Circuits Syst. **69**, 1466 (2022).
[136]J. Saikia, S. Yin, S. K. Cherupally, B. Zhang, J. Meng, M. Seok, and J.-S. Seo, in *2021 Design Automation and Test in Europe Conference and Exhibition (DATE)* (IEEE, 2021).
[137]Y.-D. Chih, P.-H. Lee, H. Fujiwara, Y.-C. Shih, C.-F. Lee, R. Naous, Y.-L. Chen, C.-P. Lo, C.-H. Lu, H. Mori, W.-C. Zhao, D. Sun, M. E. Sinangil, Y.-H. Chen, T.-L. Chou, K. Akarvardar, H.-J. Liao, Y. Wang, M.-F. Chang, and T.-Y. J. Chang, in *2021 IEEE International Solid-State Circuits Conference (ISSCC)* (IEEE, 2021).
[138]G. Desoli, N. Chawla, T. Boesch, M. Avodhyawasi, H. Rawat, H. Chawla, V. S. Abhijith, P. Zambotti, A. Sharma, C. Cappetta, M. Rossi, A. De Vita, and F. Girardi, in *2023 IEEE International Solid-State Circuits Conference (ISSCC)* (IEEE, 2023).
[139]J. Zhang, Z. Wang, and N. Verma, in *2016 IEEE Symposium on VLSI Circuits (VLSI-Circuits)* (IEEE, 2016).
[140]M. Kang, S. K. Gonugondla, A. Patil, and N. R. Shanbhag, IEEE J. Solid-State Circuits **53**, 642 (2018).
[141]H. Valavi, P. J. Ramadge, E. Nestler, and N. Verma, in *2018 IEEE Symposium on VLSI Circuits* (IEEE, 2018).
[142]H. Jia, H. Valavi, Y. Tang, J. Zhang, and N. Verma, in *2019 IEEE Hot Chips 31 Symposium (HCS)* (IEEE, 2019).
[143]H. Jia, M. Ozatay, Y. Tang, H. Valavi, R. Pathak, J. Lee, and N. Verma, in *2021 IEEE International Solid-State Circuits Conference (ISSCC)* (IEEE, 2021).
[144]K. Ueyoshi, I. A. Papistas, P. Houshmand, G. M. Sarda, V. Jain, M. Shi, Q. Zheng, S. Giraldo, P. Vrancx, J. Doevenspeck, D. Bhattacharjee, S. Cosemans, A. Mallik, P. Debacker, D. Verkest, and M. Verhelst, in *2022 IEEE International Solid-State Circuits Conference (ISSCC)* (IEEE, 2022).
[145]R. Bez, E. Camerlenghi, A. Modelli, and A. Visconti, Proc. IEEE **91**, 489 (2003).
[146]C. Monzio Compagnoni, A. Goda, A. S. Spinelli, P. Feeley, A. L. Lacaita, and A. Visconti, Proc. IEEE **105**, 1609 (2017).
[147]T. Pekny, L. Vu, J. Tsai, D. Srinivasan, E. Yu, J. Pabustan, J. Xu, S. Deshmukh, K.-F. Chan, M. Piccardi, K. Xu, G. Wang, K. Shakeri, V. Patel, T. Iwasaki, T. Wang, P. Musunuri, C. Gu, A. Mohammadzadeh, A. Ghalam, V. Moschiano, T. Vali, J. Park, J. Lee, and R. Ghodsi, in *IEEE International Solid-State Circuits Conference (ISSCC). Digest of Technical Papers* (IEEE, 2022), Vol. 132.
[148]F. Merrikh-Bayat, X. Guo, M. Klachko, M. Prezioso, K. K. Likharev, and D. B. Strukov, IEEE Trans. Neural Networks Learn. Syst. **29**, 4782 (2018).
[149]X. Guo, F. Merrikh-Bayat, M. Prezioso, Y. Chen, B. Nguyen, N. Do, and D. B. Strukov, "Temperature-insensitive analog vector-by-matrix multiplier based on 55 nm NOR flash memory cells," in *2017 IEEE Custom Integrated Circuits Conference (CICC)* (IEEE, Austin, TX, 2017), pp. 1–4.
[150]M. Bavandpour, S. Sahay, M. R. Mahmoodi, and D. B. Strukov, Neuromorphic Comput. Eng. **1**, 014001 (2021).
[151]P. Wang, F. Xu, B. Wang, B. Gao, H. Wu, H. Qian, and S. Yu, IEEE Trans. Very Large Scale Integr. Syst. **27**, 988 (2019).
[152]W. Shim and S. Yu, IEEE Electron Device Lett. **42**, 160 (2021).
[153]S.-T. Lee, G. Yeom, H. Yoo, H.-S. Kim, S. Lim, J.-H. Bae, B.-G. Park, and J.-H. Lee, IEEE Trans. Electron Devices **68**, 3365 (2021).
[154]D. Lee, Y. I. Lee *et al.*, Front. Neurosci. **14**, 14 (2020).
[155]H.-T. Lue, P.-K. Hsu, M.-L. Wei, T.-H. Yeh, P.-Y. Du, W.-C. Chen, K.-C. Wang, and C.-Y. Lu, "Optimal design methods to transform 3D NAND flash into a high-density, high-bandwidth and low-power nonvolatile computing in memory (nvCIM) accelerator for deep-learning neural networks (DNN)," in *2019 IEEE International Electron Devices Meeting (IEDM)* (IEEE, San Francisco, CA, USA, 2019), p. 38.1.1-38.1.4.
[156]H.-N. Yoo, J.-W. Back, N.-H. Kim, D. Kwon, B.-G. Park, and J.-H. Lee, in *Proceedings of IEEE Symposium on VLSI Technology and Circuits* (IEEE, 2022), Vol. 304.
[157]G. Malavena, M. Filippi, A. S. Spinelli, and C. Monzio Compagnoni, "Unsupervised learning by spike-timing-dependent plasticity in a mainstream NOR flash memory array—Part I: Cell operation," IEEE Trans. Electron Devices **66**, 4727–4732 (2019).
[158]G. Malavena, M. Filippi, A. S. Spinelli, and C. Monzio Compagnoni, IEEE Trans. Electron Devices **66**, 4733 (2019).
[159]C. Monzio Compagnoni and A. S. Spinelli, IEEE Trans. Electron Devices **66**, 4504 (2019).
[160]G. Malavena, S. Petrò, A. S. Spinelli, and C. M. Compagnoni, in *Proceedings of European Solid-State Electron Device Conference* (IEEE, 2019), p. 122.
[161]D.-H. Kwon, K. M. Kim, J. H. Jang, J. M. Jeon, M. H. Lee, G. H. Kim, X.-S. Li, G.-S. Park, B. Lee, S. Han, M. Kim, and C. S. Hwang, Nat. Nanotechnol. **5**, 148 (2010).

[162]L. Song, Y. Zhuo, X. Qian, H. Li, and Y. Chen, in *2018 IEEE International Symposium on High Performance Computer Architecture (HPCA)* (IEEE, 2018).
[163]G. Dai, T. Huang, Y. Wang, H. Yang, and J. Wawrzynek, in *Proceedings of the 24th Asia and South Pacific Design Automation Conference* (ACM, 2019).
[164]S. Wang, Y. Li, D. Wang, W. Zhang, X. Chen, D. Dong, S. Wang, X. Zhang, P. Lin, C. Gallicchio, X. Xu, Q. Liu, K.-T. Cheng, Z. Wang, D. Shang, and M. Liu, Nat. Mach. Intell. **5**, 104 (2023).
[165]W. Zhang, S. Wang, Y. Li, X. Xu, D. Dong, N. Jiang, F. Wang, Z. Guo, R. Fang, C. Dou, K. Ni, Z. Wang, D. Shang, and M. Liu, in *2022 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits)* (IEEE, 2022).
[166]F. Alibart, E. Zamanidoost, and D. B. Strukov, Nat. Commun. **4**, 2072 (2013).
[167]F. M. Bayat, M. Prezioso, B. Chakrabarti, H. Nili, I. Kataeva, and D. Strukov, Nat. Commun. **9**, 2331 (2018).
[168]M. Hu, C. E. Graves, C. Li, Y. Li, N. Ge, E. Montgomery, N. Davila, H. Jiang, R. S. Williams, J. J. Yang, Q. Xia, and J. P. Strachan, Adv. Mater. **30**, 1705914 (2018).
[169]Q. Duan, Z. Jing, X. Zou, Y. Wang, K. Yang, T. Zhang, S. Wu, R. Huang, and Y. Yang, Nat. Commun. **11**, 3399 (2020).
[170]M. Ishii, U. Shin, K. Hosokawa, M. BrightSky, W. Haensch, S. Kim, S. Lewis, A. Okazaki, J. Okazawa, M. Ito, M. Rasch, W. Kim, and A. Nomura, in *2019 IEEE International Electron Devices Meeting (IEDM)* (IEEE, 2019).
[171]S. Yu, Z. Li, P.-Y. Chen, H. Wu, B. Gao, D. Wang, W. Wu, and H. Qian, in *2016 IEEE International Electron Devices Meeting (IEDM)* (IEEE, 2016).
[172]B. Yan, Q. Yang, W.-H. Chen, K.-T. Chang, J.-W. Su, C.-H. Hsu, S.-H. Li, H.-Y. Lee, S.-S. Sheu, M.-S. Ho, Q. Wu, M.-F. Chang, Y. Chen, and H. Li, in *2019 Symposium on VLSI Technology* (IEEE, 2019).
[173]C.-X. Xue, T.-Y. Huang, J.-S. Liu, T.-W. Chang, H.-Y. Kao, J.-H. Wang, T.-W. Liu, S.-Y. Wei, S.-P. Huang, W.-C. Wei, Y.-R. Chen, T.-H. Hsu, Y.-K. Chen, Y.-C. Lo, T.-H. Wen, C.-C. Lo, R.-S. Liu, C.-C. Hsieh, K.-T. Tang, and M.-F. Chang, in *2020 IEEE International Solid-State Circuits Conference-(ISSCC)* (IEEE, 2020).
[174]Y. Lu, X. Li, L. Yan, T. Zhang, Y. Yang, Z. Song, and R. Huang, in *2020 IEEE International Electron Devices Meeting (IEDM)* (IEEE, 2020).
[175]V. Joshi, M. Le Gallo, S. Haefeli, I. Boybat, S. R. Nandakumar, C. Piveteau, M. Dazzi, B. Rajendran, A. Sebastian, and E. Eleftheriou, Nat. Commun. **11**, 2473 (2020).
[176]H. Li, W.-C. Chen, A. Levy, C.-H. Wang, H. Wang, P.-H. Chen, W. Wan, W.-S. Khwa, H. Chuang, Y.-D. Chih, M.-F. Chang, H.-S. P. Wong, and P. Raina, IEEE Trans. Electron Devices **68**, 6637 (2021).
[177]P. Yao, H. Wu, B. Gao, S. B. Eryilmaz, X. Huang, W. Zhang, Q. Zhang, N. Deng, L. Shi, H.-S. P. Wong, and H. Qian, Nat. Commun. **8**, 15199 (2017).
[178]P. Yao, H. Wu, B. Gao, J. Tang, Q. Zhang, W. Zhang, J. J. Yang, and H. Qian, Nature **577**, 641 (2020).
[179]S. Ambrogio, P. Narayanan, H. Tsai, R. M. Shelby, I. Boybat, C. di Nolfo, S. Sidler, M. Giordano, M. Bodini, N. C. P. Farinha, B. Killeen, C. Cheng, Y. Jaoudi, and G. W. Burr, Nature **558**, 60 (2018).
[180]G. Karunaratne, M. Schmuck, M. Le Gallo, G. Cherubini, L. Benini, A. Sebastian, and A. Rahimi, Nat. Commun. **12**, 2468 (2021).
[181]P. M. Sheridan, F. Cai, C. Du, W. Ma, Z. Zhang, and W. D. Lu, Nat. Nanotechnol. **12**, 784 (2017).
[182]W. Wan, R. Kubendran, S. B. Eryilmaz, W. Zhang, Y. Liao, D. Wu, S. Deiss, B. Gao, P. Raina, S. Joshi, H. Wu, G. Cauwenberghs, and H.-S. P. Wong, in *2020 IEEE International Solid-State Circuits Conference (ISSCC)* (IEEE, 2020).
[183]S. Woźniak, A. Pantazi, T. Bohnstingl, and E. Eleftheriou, Nat. Mach. Intell. **2**, 325 (2020).
[184]C. Du, F. Cai, M. A. Zidan, W. Ma, S. H. Lee, and W. D. Lu, Nat. Commun. **8**, 2204 (2017).
[185]J. Moon, W. Ma, J. H. Shin, F. Cai, C. Du, S. H. Lee, and W. D. Lu, Nat. Electron. **2**, 480 (2019).
[186]H. Tsai, S. Ambrogio, C. Mackin, P. Narayanan, R. M. Shelby, K. Rocki, A. Chen, and G. W. Burr, in *2019 Symposium on VLSI Technology* (IEEE, 2019).
[187]G. Karunaratne, M. L. Gallo, G. Cherubini, L. Benini, A. Rahimi, and A. Sebastian, Nat. Electron. **3**, 327 (2020).
[188]Z. Liu, J. Tang, B. Gao, P. Yao, X. Li, D. Liu, Y. Zhou, H. Qian, B. Hong, and H. Wu, Nat. Commun. **11**, 4234 (2020).
[189]Y. Zhong, J. Tang, X. Li, B. Gao, H. Qian, and H. Wu, Nat. Commun. **12**, 408 (2021).
[190]Z. Sun, G. Pedretti, A. Bricalli, and D. Ielmini, Sci. Adv. **6**, eaay2378 (2020).
[191]Y. J. Jeong, J. Lee, J. Moon, J. H. Shin, and W. D. Lu, Nano Lett. **18**, 4447 (2018).
[192]S. Choi, J. H. Shin, J. Lee, P. Sheridan, and W. D. Lu, Nano Lett. **17**, 3113 (2017).
[193]A. Sebastian, T. Tuma, N. Papandreou, M. Le Gallo, L. Kull, T. Parnell, and E. Eleftheriou, Nat. Commun. **8**, 1115 (2017).
[194]S. Choi, S. H. Tan, Z. Li, Y. Kim, C. Choi, P.-Y. Chen, H. Yeon, S. Yu, and J. Kim, Nat. Mater. **17**, 335 (2018).
[195]M. Rao, H. Tang, J. Wu, W. Song, M. Zhang, W. Yin, Y. Zhuo, F. Kiani, B. Chen, X. Jiang, H. Liu, H.-Y. Chen, R. Midya, F. Ye, H. Jiang, Z. Wang, M. Wu, M. Hu, H. Wang, Q. Xia, N. Ge, J. Li, and J. J. Yang, "Thousands of conductance levels in memristors integrated on CMOS," Nature **615**(7954), 823–829 (2023).
[196]F. Kiani, J. Yin, Z. Wang, J. J. Yang, and Q. Xia, Sci. Adv. **7**, eabj4801 (2021).
[197]T. Dalgaty, N. Castellani, C. Turck, K.-E. Harabi, D. Querlioz, and E. Vianello, Nat. Electron. **4**, 151 (2021).
[198]Z. Wang, C. Li, P. Lin, M. Rao, Y. Nie, W. Song, Q. Qiu, Y. Li, P. Yan, J. P. Strachan, N. Ge, N. McDonald, Q. Wu, M. Hu, H. Wu, R. S. Williams, Q. Xia, and J. J. Yang, Nat. Mach. Intell. **1**, 434 (2019).
[199]M. Le Gallo, A. Sebastian, R. Mathis, M. Manica, H. Giefers, T. Tuma, C. Bekas, A. Curioni, and E. Eleftheriou, Nat. Electron. **1**, 246 (2018).
[200]F. Cai, S. Kumar, T. V. Vaerenbergh, X. Sheng, R. Liu, C. Li, Z. Liu, M. Foltin, S. Yu, Q. Xia, J. J. Yang, R. Beausoleil, W. D. Lu, and J. P. Strachan, Nat. Electron. **3**, 409 (2020).
[201]M. R. Mahmoodi, M. Prezioso, and D. B. Strukov, Nat. Commun. **10**, 5113 (2019).
[202]M. Le Gallo and A. Sebastian, J. Phys. D: Appl. Phys. **53**, 213002 (2020).
[203]S. G. Sarwat, Mater. Sci. Technol. **33**, 1890 (2017).
[204]A. Sebastian, M. Le Gallo, G. W. Burr, S. Kim, M. BrightSky, and E. Eleftheriou, J. Appl. Phys. **124**, 111101 (2018).
[205]S. R. Ovshinsky and B. Pashmakov, MRS Online Proc. Libr. **803**, 61 (2003).
[206]M. L. Gallo, R. Khaddam-Aljameh, M. Stanisavljevic, A. Vasilopoulos, B. Kersting, M. Dazzi, G. Karunaratne, M. Braendli, A. Singh, S. M. Mueller *et al.*, Nat. Electron. **6**, 680 (2023).
[207]S. R. Nandakumar, M. Le Gallo, C. Piveteau, V. Joshi, G. Mariani, I. Boybat, G. Karunaratne, R. Khaddam-Aljameh, U. Egger, A. Petropoulos *et al.*, Front. Neurosci. **14**, 406 (2020).
[208]G. Karunaratne, M. Hersche, J. Langeneager, G. Cherubini, M. Le Gallo, U. Egger, K. Brew, S. Choi, I. Ok, C. Silvestre *et al.*, in *ESSCIRC 2022-IEEE 48th European Solid State Circuits Conference (ESSCIRC)* (IEEE, 2022), pp. 105–108.
[209]S. Ambrogio, N. Ciocchini, M. Laudato, V. Milo, A. Pirovano, P. Fantini, and D. Ielmini, Front. Neurosci. **10**, 56 (2016).
[210]T. Tuma, M. Le Gallo, A. Sebastian, and E. Eleftheriou, IEEE Electron Device Lett. **37**, 1238 (2016).
[211]T. Tuma, A. Pantazi, M. Le Gallo, A. Sebastian, and E. Eleftheriou, Nat. Nanotechnol. **11**, 693 (2016).
[212]See https://download.intel.com/newsroom/2021/archive/2015-07-28-news-releases-intel-and-micron-produce-breakthrough-memory-technology.pdf for Intel report.
[213]F. Arnaud, P. Ferreira, F. Piazza, A. Gandolfo, P. Zuliani, P. Mattavelli, E. Gomiero, G. Samanni, J. Jasse, C. Jahan *et al.*, in *International Electron Devices Meeting (IEDM)* (IEEE, 2020), pp. 2–24.
[214]M. Lanza, A. Sebastian, W. D. Lu, M. Le Gallo, M.-F. Chang, D. Akinwande, F. M. Puglisi, H. N. Alshareef, M. Liu, and J. B. Roldan, Science **376**, 9979 (2022).
[215]S. Raoux, Annu. Rev. Mater. Res. **39**, 25 (2009).
[216]V. Joshi, M. Le Gallo, S. Haefeli, I. Boybat, S. R. Nandakumar, C. Piveteau, M. Dazzi, B. Rajendran, A. Sebastian, and E. Eleftheriou, Nat. Commun. **11**, 2473 (2020).
[217]I. Boybat, B. Kersting, S. G. Sarwat, X. Timoneda, R. L. Bruce, M. BrightSky, M. Le Gallo, and A. Sebastian, in *2021 IEEE International Electron Devices Meeting (IEDM)* (IEEE, 2021), pp. 28–33.
[218]S. G. Sarwat, M. Le Gallo, R. L. Bruce, K. Brew, B. Kersting, V. P. Jonnalagadda, I. Ok, N. Saulnier, M. BrightSky, and A. Sebastian, Adv. Mater. **35**, 2201238 (2022).
[219]S. R. Nandakumar, I. Boybat, M. Le Gallo, E. Eleftheriou, A. Sebastian, and B. Rajendran, Sci. Rep. **10**, 1 (2020).
[220]F. Xiong, A. D. Liao, D. Estrada, and E. Pop, Science **332**, 568 (2011).

[221]Z. Yang, B. Li, J.-J. Wang, X.-D. Wang, M. Xu, H. Tong, X. Cheng, L. Lu, C. Jia, M. Xu *et al.*, Adv. Sci. **9**, 2103478 (2022).
[222]X. Wang, S. Song, H. Wang, T. Guo, Y. Xue, R. Wang, H. S. Wang, L. Chen, C. Jiang, C. Chen *et al.*, Adv. Sci. **9**, 2202222 (2022).
[223]S. G. Sarwat, P. Gehring, G. Rodriguez Hernandez, J. H. Warner, G. A. D. Briggs, J. A. Mol, and H. Bhaskaran, Nano. Lett. **17**, 3688 (2017).
[224]T. C. Chong, L. P. Shi, W. Qiang, P. K. Tan, X. S. Miao, and X. Hu, J. Appl. Phys. **91**, 3981 (2002).
[225]R. E. Simpson, P. Fons, A. V. Kolobov, T. Fukaya, M. Krbal, T. Yagi, and J. Tominaga, Nat. Nanotechnol. **6**, 501 (2011).
[226]K. L. Okabe, A. Sood, E. Yalon, C. M. Neumann, M. Asheghi, E. Pop, K. E. Goodson, and H.-S. P. Wong, J. Appl. Phys. **125**, 184501 (2019).
[227]A. I. Khan, H. Kwon, M. E. Chen, M. Asheghi, H.-S. P. Wong, K. E. Goodson, and E. Pop, IEEE Electron. Device Lett. **43**, 204 (2021).
[228]A. I. Khan, C. Perez, X. Wu, B. Won, K. Kim, H. Kwon, P. Ramesh, K. M. Neilson, M. Asheghi, K. Saraswat *et al.*, in *2022 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits)* (IEEE, 2022), pp. 310–311.
[229]W. Zhang, R. Mazzarello, M. Wuttig, and E. Ma, Nat. Rev. Mater. **4**, 150 (2019).
[230]A. Sebastian, M. Le Gallo, and D. Krebs, Nat. Commun. **5**, 4314 (2014).
[231]F. Rao, K. Ding, Y. Zhou, Y. Zheng, M. Xia, S. Lv, Z. Song, S. Feng, I. Ronneberger, R. Mazzarello *et al.*, Science **358**, 1423 (2017).
[232]Y. Xie, W. Kim, Y. Kim, S. Kim, J. Gonsalves, M. BrightSky, C. Lam, Y. Zhu, and J. J. Cha, Adv. Mater. **30**, 1705587 (2018).
[233]W. W. Koelmans, A. Sebastian, V. P. Jonnalagadda, D. Krebs, L. Dellmann, and E. Eleftheriou, Nat. Commun. **6**, 8181 (2015).
[234]S. G. Sarwat, T. M. Philip, C.-T. Chen, B. Kersting, R. L. Bruce, C.-W. Cheng, N. Li, N. Saulnier, M. BrightSky, and A. Sebastian, Adv. Funct. Mater. **31**, 2106547 (2021).
[235]M. Salinga, B. Kersting, I. Ronneberger, V. P. Jonnalagadda, X. T. Vu, M. Le Gallo, I. Giannopoulos, O. Cojocaru-Mirédin, R. Mazzarello, and A. Sebastian, Nat. Mater. **17**, 681 (2018).
[236]I. Boybat, M. Le Gallo, S. R. Nandakumar, T. Moraitis, T. Parnell, T. Tuma, B. Rajendran, Y. Leblebici, A. Sebastian, and E. Eleftheriou, Nat. Commun. **9**, 2514 (2018).
[237]J. Büchel, A. Vasilopoulos, B. Kersting, F. Odermatt, K. Brew, I. Ok, S. Choi, I. Saraf, V. Chan, T. Philip *et al.*, in *International Electron Devices Meeting (IEDM)* (IEEE, 2022), p. 33–1.
[238]C. Ríos, M. Stegmaier, P. Hosseini, D. Wang, T. Scherer, C. D. Wright, H. Bhaskaran, and W. H. P. Pernice, Nat. Photon. **9**, 725 (2015).
[239]C. Ríos, N. Youngblood, Z. Cheng, M. Le Gallo, W. H. P. Pernice, C. D. Wright, A. Sebastian, and H. Bhaskaran, Sci. Adv. **5**, 5759 (2019).
[240]B. J. Shastri, A. N. Tait, T. Ferreira de Lima, W. H. P. Pernice, H. Bhaskaran, C. D. Wright, and P. R. Prucnal, Nat. Photon. **15**, 102 (2021).
[241]J. Feldmann, N. Youngblood, M. Karpov, H. Gehring, X. Li, M. Stappers, M. Le Gallo, X. Fu, A. Lukashchuk, A. S. Raja *et al.*, Nature **589**, 52 (2021).
[242]C. Wu, H. Yu, S. Lee, R. Peng, I. Takeuchi, and M. Li, Nat. Commun. **12**, 96 (2021).
[243]S. G. Sarwat, F. Brückerhoff-Plückelmann, S. G.-C. Carrillo, E. Gemo, J. Feldmann, H. Bhaskaran, C. D. Wright, W. H. P. Pernice, and A. Sebastian, Sci. Adv. **8**, 3243 (2022).
[244]Y. Lu, X. Li, B. Yan, L. Yan, T. Zhang, Z. Song, R. Huang, and Y. Yang, Adv. Mater. **34**, 2107811 (2022).
[245]J. Langenegger, G. Karunaratne, M. Hersche, L. Benini, A. Sebastian, and A. Rahimi, "In-memory factorization of holographic perceptual representations," Nat. Nanotechnol. **18**(5), 479–485 (2023).
[246]S. G. Sarwat, B. Kersting, T. Moraitis, V. P. Jonnalagadda, and A. Sebastian, Nat. Nanotechnol. **17**, 507 (2022).
[247]J. Liang, S. Yeh, S. S. Wong, and H.-S. P. Wong, in *4th IEEE International Memory Workshop* (IEEE, 2012), Vol. 1.
[248]S. Yoo, H. D. Lee, S. Lee, H. Choi, and T. Kim, IEEE Trans. Electron Devices **67**, 1454 (2020).
[249]S. H. Lee, M. S. Kim, G. S. Do, S. G. Kim, H. J. Lee, J. S. Sim, N. G. Park, S. B. Hong, Y. H. Jeon, K. S. Choi *et al.*, in *2010 Symposium on VLSI Technology* (IEEE, 2010), pp. 199–200.
[250]T. Mikolajick, M. H. Park, L. Begon-Lours, and S. Slesazeck, Adv. Mater. **35**, 2206042 (2022).
[251]H. P. McAdams, R. Acklin, T. Blake, X.-H. Du, J. Eliason, J. Fong, W. F. Kraus, D. Liu, S. Madan, T. Moise, S. Natarajan, N. Qian, Y. Qiu, K. A. Remack, J. Rodriguez, J. Roscher, A. Seshadri, and S. R. Summerfelt, IEEE J. Solid-State Circuits **39**, 667 (2004).
[252]S. Beyer, S. Dunkel, M. Trentzsch, J. Muller, A. Hellmich, D. Utess, J. Paul, D. Kleimaier, J. Pellerin, S. Muller, J. Ocker, A. Benoist, H. Zhou, M. Mennenga, M. Schuster, F. Tassan, M. Noack, A. Pourkeramati, F. Muller, M. Lederer, T. Ali, R. Hoffmann, T. Kampfe, K. Seidel, H. Mulaosmanovic, E. T. Breyer, T. Mikolajick, and S. Slesazeck, in *2020 IEEE International Memory Workshop (IMW)* (IEEE, 2020).
[253]E. Y. Tsymbal and H. Kohlstedt, Science **313**, 181 (2006).
[254]K. Asadi, M. Li, P. W. M. Blom, M. Kemerink, and D. M. de Leeuw, Mater. Today **14**, 592 (2011).
[255]D. Bondurant, Ferroelectrics **112**, 273 (1990).
[256]T. S. Böscke, J. Müller, D. Bräuhaus, U. Schröder, and U. Böttger, Appl. Phys. Lett. **99**, 102903 (2011).
[257]H. Mulaosmanovic, P. D. Lomenzo, U. Schroeder, S. Slesazeck, T. Mikolajick, and B. Max, in *2021 IEEE International Reliability Physics Symposium (IRPS)* (IEEE, 2021).
[258]R. Materlik, C. Künneth, and A. Kersch, J. Appl. Phys. **117**, 134109 (2015).
[259]H. Mulaosmanovic, S. Dunkel, M. Trentzsch, S. Beyer, E. T. Breyer, T. Mikolajick, and S. Slesazeck, IEEE Trans. Electron Devices **67**, 5804 (2020).
[260]M. Materano, P. D. Lomenzo, A. Kersch, M. H. Park, T. Mikolajick, and U. Schroeder, Inorg. Chem. Front. **8**, 2650 (2021).
[261]D. R. Islamov, V. A. Gritsenko, T. V. Perevalov, V. A. Pustovarov, O. M. Orlov, A. G. Chernikova, A. M. Markeev, S. Slesazeck, U. Schroeder, T. Mikolajick, and G. Y. Krasnikov, Acta Mater. **166**, 47 (2019).
[262]L. Bégon-Lours, M. Halter, F. M. Puglisi, L. Benatti, D. F. Falcone, Y. Popoff, D. Dávila Pineda, M. Sousa, and B. J. Offrein, Adv. Electron. Mater. **8**, 2101395 (2022).
[263]P. Nukala, M. Ahmadi, Y. Wei, S. de Graaf, E. Stylianidis, T. Chakrabortty, S. Matzen, H. W. Zandbergen, A. Björling, D. Mannix, D. Carbone, B. Kooi, and B. Noheda, Science **372**, 630 (2021).
[264]S. S. Cheema, D. Kwon, N. Shanker, R. dos Reis, S.-L. Hsu, J. Xiao, H. Zhang, R. Wagner, A. Datar, M. R. McCarter, C. R. Serrao, A. K. Yadav, G. Karbasian, C.-H. Hsu, A. J. Tan, L.-C. Wang, V. Thakare, X. Zhang, A. Mehta, E. Karapetrova, R. V. Chopdekar, P. Shafer, E. Arenholz, C. Hu, R. Proksch, R. Ramesh, J. Ciston, and S. Salahuddin, Nature **580**, 478 (2020).
[265]Y. Wei, G. Vats, and B. Noheda, Neuromorphic Comput. Eng. **2**, 044007 (2022).
[266]B. Dieny, I. L. Prejbeanu, K. Garello, P. Gambardella, P. Freitas, R. Lehndorff, W. Raberg, U. Ebels, S. O. Demokritov, J. Akerman, A. Deac, P. Pirro, C. Adelmann, A. Anane, A. V. Chumak, A. Hirohata, S. Mangin, S. O. Valenzuela, M. C. Onbaşlı, M. d'Aquino, G. Prenat, G. Finocchio, L. Lopez-Diaz, R. Chantrell, O. Chubykalo-Fesenko, and P. Bortolotti, Nat. Electron. **3**, 446 (2020).
[267]D. Apalkov, B. Dieny, and J. M. Slaughter, Proc. IEEE **104**, 1796 (2016).
[268]S. Lequeux, J. Sampaio, V. Cros, K. Yakushiji, A. Fukushima, R. Matsumoto, H. Kubota, S. Yuasa, and J. Grollier, Sci. Rep. **6**, 31510 (2016).
[269]S. Fukami, C. Zhang, S. DuttaGupta, A. Kurenkov, and H. Ohno, Nat. Mater. **15**, 535 (2016).
[270]J. Wang, H. Sepehri-Amin, H. Tajiri, T. Nakamura, K. Masuda, Y. K. Takahashi, T. Ina, T. Uruga, I. Suzuki, Y. Miura, and K. Hono, Acta Mater. **166**, 413 (2019).
[271]K. M. Song, J.-S. Jeong, B. Pan, X. Zhang, J. Xia, S. Cha, T.-E. Park, K. Kim, S. Finizio, J. Raabe, J. Chang, Y. Zhou, W. Zhao, W. Kang, H. Ju, and S. Woo, Nat. Electron. **3**, 148 (2020).
[272]E. Raymenants, A. Vaysset, D. Wan, M. Manfrini, O. Zografos, O. Bultynck, J. Doevenspeck, M. Heyns, I. P. Radu, and T. Devolder, J. Appl. Phys. **124**, 152116 (2018).
[273]M. Prezioso, M. R. Mahmoodi, F. M. Bayat, H. Nili, H. Kim, A. Vincent, and D. B. Strukov, Nat. Commun. **9**, 5311 (2018).
[274]F.-X. Liang, I.-T. Wang, and T.-H. Hou, Adv. Intell. Syst. **3**, 2100007 (2021).

[275]J. Torrejon, M. Riou, F. A. Araujo, S. Tsunegi, G. Khalsa, D. Querlioz, P. Bortolotti, V. Cros, K. Yakushiji, A. Fukushima, H. Kubota, S. Yuasa, M. D. Stiles, and J. Grollier, Nature **547**, 428 (2017).
[276]K. Hayakawa, S. Kanai, T. Funatsu, J. Igarashi, B. Jinnai, W. A. Borders, H. Ohno, and S. Fukami, Phys. Rev. Lett. **126**, 117202 (2021).
[277]W. A. Borders, A. Z. Pervaiz, S. Fukami, K. Y. Camsari, H. Ohno, and S. Datta, Nature **573**, 390 (2019).
[278]M.-H. Wu, M.-C. Hong, C.-C. Chang, P. Sahu, J.-H. Wei, H.-Y. Lee, S.-S. Shcu, and T.-H. Hou, in *2019 Symposium on VLSI Technology* (IEEE, 2019).
[279]M.-H. Wu, I.-T. Wang, M.-C. Hong, K.-M. Chen, Y.-C. Tseng, J.-H. Wei, and T.-H. Hou, Phys. Rev. Appl. **18**, 064034 (2022).
[280]M. Schott, A. Bernand-Mantel, L. Ranno, S. Pizzini, J. Vogel, H. Béa, C. Baraduc, S. Auffret, G. Gaudin, and D. Givord, Nano Lett. **17**, 3006 (2017).
[281]C. E. Fillion, J. Fischer, R. Kumar, A. Fassatoui, S. Pizzini, L. Ranno, D. Ourdani, M. Belmeguenai, Y. Roussigné, S. M. Chérif, S. Auffret, I. Joumard, O. Boulle, G. Gaudin, L. Buda-Prejbeanu, C. Baraduc, and H. Béa, Nat. Commun. **13**, 5257 (2022).
[282]A. Jaiswal, S. Roy, G. Srinivasan, and K. Roy, IEEE Trans. Electron Devices **64**, 1818 (2017).
[283]M. Zahedinejad, H. Fulara, R. Khymyn, A. Houshang, M. Dvornik, S. Fukami, S. Kanai, H. Ohno, and J. Åkerman, Nat. Mater. **21**, 81 (2022).
[284]X.-G. Zhang and W. H. Butler, Phys. Rev. B **70**, 172407 (2004).
[285]A. Ross, N. Leroux, A. De Riz, D. Marković, D. Sanz-Hernández, J. Trastoy, P. Bortolotti, D. Querlioz, L. Martins, L. Benetti, M. S. Claro, P. Anacleto, A. Schulman, T. Taris, J.-B. Begueret, S. Saïghi, A. S. Jenkins, R. Ferreira, A. F. Vincent, F. A. Mizrahi, and J. Grollier, Nat. Nanotechnol. **18**, 1273 (2023).
[286]S. Kanai, K. Hayakawa, H. Ohno, and S. Fukami, Phys. Rev. B **103**, 094423 (2021).
[287]M.-H. Wu, M.-S. Huang, Z. Zhu, F.-X. Liang, M.-C. Hong, J. Deng, J.-H. Wei, S.-S. Sheu, C.-I. Wu, G. Liang, and T.-H. Hou, in *2020 IEEE Symposium on VLSI Technology* (IEEE, 2020).
[288]J. Choe, in *2023 IEEE International Memory Workshop (IMW)* (IEEE, 2023).
[289]R. Khymyn, I. Lisenkov, J. Voorheis, O. Sulymenko, O. Prokopenko, V. Tiberkevich, J. Akerman, and A. Slavin, Sci. Rep. **8**, 15727 (2018).
[290]F. Trier, P. Noël, J.-V. Kim, J.-P. Attané, L. Vila, and M. Bibes, Nat. Rev. Mater. **7**, 258 (2021).
[291]S. S. P. Parkin, M. Hayashi, and L. Thomas, Science **320**, 190 (2008).
[292]A. Fert, V. Cros, and J. Sampaio, Nat. Nanotechnol. **8**, 152 (2013).
[293]M. S. El Hadri, P. Pirro, C.-H. Lambert, S. Petit-Watelot, Y. Quessab, M. Hehn, F. Montaigne, G. Malinowski, and S. Mangin, Phys. Rev. B **94**, 064412 (2016).
[294]A. Fernández-Pacheco, R. Streubel, O. Fruchart, R. Hertel, P. Fischer, and R. P. Cowburn, Nat. Commun. **8**, 15756 (2017).
[295]Y. Shen, N. C. Harris, S. Skirlo, M. Prabhu, T. Baehr-Jones, M. Hochberg, X. Sun, S. Zhao, H. Larochelle, D. Englund, and M. Soljačić, Nat. Photonics **11**, 441 (2017).
[296]S. Bandyopadhyay, A. Sludds, S. Krastanov, R. Hamerly, N. Harris, D. Bunandar, M. Streshinsky, M. Hochberg, and D. Englund, arXiv:2208.01623 [Cs.ET] (2022).
[297]J. Feldmann, N. Youngblood, C. D. Wright, H. Bhaskaran, and W. H. P. Pernice, Nature **569**, 208 (2019).
[298]J. Zhang, S. Dai, Y. Zhao, J. Zhang, and J. Huang, Adv. Intell. Syst. **2**, 1900136 (2020).
[299]M. Kumar, S. Abbas, and J. Kim, ACS Appl. Mater. Interfaces **10**, 34370 (2018).
[300]L. Hu, J. Yang, J. Wang, P. Cheng, L. O. Chua, and F. Zhuge, Adv. Funct. Mater. **31**, 2005582 (2020).
[301]C. Lu, J. Meng, J. Song, T. Wang, H. Zhu, Q.-Q. Sun, D. W. Zhang, and L. Chen, Nano Lett. **24**, 1667 (2024).
[302]I. Taghavi, M. Moridsadat, A. Tofini, S. Raza, N. A. F. Jaeger, L. Chrostowski, B. J. Shastri, and S. Shekhar, Nanophotonics **11**, 3855 (2022).
[303]J. Wu, Z. T. Xie, Y. Sha, H. Y. Fu, and Q. Li, Photonics Res. **9**, 1616 (2021).
[304]J. Robertson, P. Kirkland, J. A. Alanis, M. Hejda, J. Bueno, G. Di Caterina, and A. Hurtado, Sci. Rep. **12**, 4874 (2022).
[305]S. T. Ilie, J. Faneca, I. Zeimpekis, T. D. Bucio, K. Grabska, D. W. Hewak, H. M. H. Chong, and F. Y. Gardes, Sci. Rep. **12**, 17815 (2022).
[306]M. Wang, S. Cai, C. Pan, C. Wang, X. Lian, Y. Zhuo, K. Xu, T. Cao, X. Pan, B. Wang, S.-J. Liang, J. J. Yang, P. Wang, and F. Miao, Nat. Electron. **1**, 130 (2018).
[307]Y. Shi, X. Liang, B. Yuan, V. Chen, H. Li, F. Hui, Z. Yu, F. Yuan, E. Pop, H.-S. P. Wong, and M. Lanza, Nat. Electron. **1**, 458 (2018).
[308]L. Sun, Y. Zhang, G. Han, G. Hwang, J. Jiang, B. Joo, K. Watanabe, T. Taniguchi, Y.-M. Kim, W. J. Yu, B.-S. Kong, R. Zhao, and H. Yang, Nat. Commun. **10**, 3161 (2019).
[309]F. M. Puglisi, L. Larcher, C. Pan, N. Xiao, Y. Shi, F. Hui, and M. Lanza, in *2016 IEEE International Electron Devices Meeting (IEDM)* (IEEE, 2016).
[310]X. Yan, J. H. Qian, V. K. Sangwan, and M. C. Hersam, Adv. Mater. **34**, 2108025 (2022).
[311]See https://www.fujitsu.com/jp/group/fsm/en/products/reram for the Fujitsu ReRAM technology specification sheet.
[312]S. Chen, M. R. Mahmoodi, Y. Shi, C. Mahata, B. Yuan, X. Liang, C. Wen, F. Hui, D. Akinwande, D. B. Strukov, and M. Lanza, Nat. Electron. **3**, 638 (2020).
[313]K. Lu, X. Li, Q. Sun, X. Pang, J. Chen, T. Minari, X. Liu, and Y. Song, Mater. Horiz. **8**, 447 (2021).
[314]S. Pazos, X. Xu, T. Guo, K. Zhu, H. N. Alshareef, and M. Lanza, Nat. Rev. Mater. **9**, 358 (2024).
[315]F. Xue, C. Zhang, Y. Ma, Y. Wen, X. He, B. Yu, and X. Zhang, Adv. Mater. **34**, 2201880 (2022).
[316]C. Kaspar, B. J. Ravoo, W. G. van der Wiel, S. V. Wegner, and W. H. P. Pernice, Nature **594**, 345 (2021).
[317]A. Chen, in *71st Device Research Conference* (IEEE, 2013).
[318]Y. Shen, W. Zheng, K. Zhu, Y. Xiao, C. Wen, Y. Liu, X. Jing, and M. Lanza, Adv. Mater. **33**, 2103656 (2021).
[319]B. Tang, H. Veluri, Y. Li, Z. G. Yu, M. Waqar, J. F. Leong, M. Sivan, E. Zamburg, Y.-W. Zhang, J. Wang, and A. V.-Y. Thean, Nat. Commun. **13**, 3037 (2022).
[320]X. Xu, T. Guo, H. Kim, M. K. Hota, R. S. Alsaadi, M. Lanza, X. Zhang, and H. N. Alshareef, Adv. Mater. **34**, 2108258 (2022).
[321]T.-A. Chen, C.-P. Chuu, C.-C. Tseng, C.-K. Wen, H.-S. P. Wong, S. Pan, R. Li, T.-A. Chao, W.-C. Chueh, Y. Zhang, Q. Fu, B. I. Yakobson, W.-H. Chang, and L.-J. Li, Nature **579**, 219 (2020).
[322]K. Y. Ma, L. Zhang, S. Jin, Y. Wang, S. I. Yoon, H. Hwang, J. Oh, D. S. Jeong, M. Wang, S. Chatterjee, G. Kim, A.-R. Jang, J. Yang, S. Ryu, H. Y. Jeong, R. S. Ruoff, M. Chhowalla, F. Ding, and H. S. Shin, Nature **606**, 88 (2022).
[323]G. Migliato Marega, H. G. Ji, Z. Wang, G. Pasquale, M. Tripathi, A. Radenovic, and A. Kis, Nat. Electron. **6**, 991 (2023).
[324]M. Lanza, R. Waser, D. Ielmini, J. J. Yang, L. Goux, J. Suñe, A. J. Kenyon, A. Mehonic, S. Spiga, V. Rana, S. Wiefels, S. Menzel, I. Valov, M. A. Villena, E. Miranda, X. Jing, F. Campabadal, M. B. Gonzalez, F. Aguirre, F. Palumbo, K. Zhu, J. B. Roldan, F. M. Puglisi, L. Larcher, T.-H. Hou, T. Prodromakis, Y. Yang, P. Huang, T. Wan, Y. Chai, K. L. Pey, N. Raghavan, S. Dueñas, T. Wang, Q. Xia, and S. Pazos, ACS Nano **15**, 17214 (2021).
[325]D. Edelstein, M. Rizzolo, D. Sil, A. Dutta, J. DeBrosse, M. Wordeman, A. Arceo, I. C. Chu, J. Demarest, E. R. J. Edwards, E. R. Evarts, J. Fullam, A. Gasasira, G. Hu, M. Iwatake, R. Johnson, V. Katragadda, T. Levin, J. Li, Y. Liu, C. Long, T. Maffitt, S. McDermott, S. Mehta, V. Mehta, D. Metzler, J. Morillo, Y. Nakamura, S. Nguyen, P. Nieves, V. Pai, R. Patlolla, R. Pujari, R. Southwick, T. Standaert, O. van der Straten, H. Wu, C.-C. Yang, D. Houssameddine, J. M. Slaughter, and D. C. Worledge, in *2020 IEEE International Electron Devices Meeting (IEDM)* (IEEE, 2020).
[326]F. Xiong, E. Yalon, A. Behnam, C. M. Neumann, K. L. Grosse, S. Deshmukh, and E. Pop, in *2016 IEEE International Electron Devices Meeting (IEDM)* (IEEE, 2016).
[327]R. Dittmann, S. Menzel, and R. Waser, Adv. Phys. **70**, 155 (2021).
[328]V. V. Zhirnov, R. Meade, R. K. Cavin, and G. Sandhu, Nanotechnology **22**, 254027 (2011).
[329]E. Chicca and G. Indiveri, Appl. Phys. Lett. **116**, 120501 (2020).
[330]C. Safranski, G. Hu, J. Z. Sun, P. Hashemi, S. L. Brown, L. Buzi, C. P. D'Emic, E. R. J. Edwards, E. Galligan, M. G. Gottwald, O. Gunawan, S. Karimeddiny, H. Jung, J. Kim, K. Latzko, P. L. Trouilloud, S. Zare, and D. C. Worledge, in *2022 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits)* (IEEE, 2022).

[331]C. Persch, M. J. Müller, A. Yadav, J. Pries, N. Honné, P. Kerres, S. Wei, H. Tanaka, P. Fantini, E. Varesi, F. Pellizzer, and M. Wuttig, Nat. Commun. **12**, 4978 (2021).
[332]M. von Witzleben, S. Wiefels, A. Kindsmüller, P. Stasner, F. Berg, F. Cüppers, S. Hoffmann-Eifert, R. Waser, S. Menzel, and U. Böttger, ACS Appl. Electron. Mater. **3**, 5563 (2021).
[333]S. Wiefels, M. V. Witzleben, M. Huttemann, U. Bottger, R. Waser, and S. Menzel, IEEE Trans. Electron Devices **68**, 1024 (2021).
[334]N. Kopperberg, S. Wiefels, S. Liberda, R. Waser, and S. Menzel, ACS Appl. Mater. Interfaces **13**, 58066 (2021).
[335]C. Bengel, J. Mohr, S. Wiefels, A. Singh, A. Gebregiorgis, R. Bishnoi, S. Hamdioui, R. Waser, D. Wouters, and S. Menzel, Neuromorphic Comput. Eng. **2**, 034001 (2022).
[336]K. Schnieders, C. Funck, F. Cüppers, S. Aussen, T. Kempen, A. Sarantopoulos, R. Dittmann, S. Menzel, V. Rana, S. Hoffmann-Eifert, and S. Wiefels, APL Mater. **10**, 101114 (2022).
[337]S. Slesazeck, V. Havel, E. Breyer, H. Mulaosmanovic, M. Hoffmann, B. Max, S. Duenkel, and T. Mikolajick, in *2019 IEEE 11th International Memory Workshop (IMW)* (IEEE, 2019).
[338]J. H. Yoon, S. J. Song, I. H. Yoo, J. Y. Seok, K. J. Yoon, D. E. Kwon, T. H. Park, and C. S. Hwang, Adv. Funct. Mater. **24**, 5086 (2014).
[339]M. Lanza, H.-S. P. Wong, E. Pop, D. Ielmini, D. Strukov, B. C. Regan, L. Larcher, M. A. Villena, J. J. Yang, L. Goux, A. Belmonte, Y. Yang, F. M. Puglisi, J. Kang, B. Magyari-Köpe, E. Yalon, A. Kenyon, M. Buckwell, A. Mehonic, A. Shluger, H. Li, T.-H. Hou, B. Hudec, D. Akinwande, R. Ge, S. Ambrogio, J. B. Roldan, E. Miranda, J. Suñe, K. L. Pey, X. Wu, N. Raghavan, E. Wu, W. D. Lu, G. Navarro, W. Zhang, H. Wu, R. Li, A. Holleitner, U. Wurstbauer, M. C. Lemme, M. Liu, S. Long, Q. Liu, H. Lv, A. Padovani, P. Pavan, I. Valov, X. Jing, T. Han, K. Zhu, S. Chen, F. Hui, and Y. Shi, Adv. Electron. Mater. **5**, 1800143 (2018).
[340]G. Dearnaley, A. M. Stoneham, and D. V. Morgan, Rep. Prog. Phys. **33**, 1129 (1970).
[341]Y. Yang, P. Gao, S. Gaba, T. Chang, X. Pan, and W. Lu, Nat. Commun. **3**, 732 (2012).
[342]Y. Yang, P. Gao, L. Li, X. Pan, S. Tappertzhofen, S. H. Choi, R. Waser, I. Valov, and W. D. Lu, Nat. Commun. **5**, 4232 (2014).
[343]R. Waser, R. Dittmann, G. Staikov, and K. Szot, Adv. Mater. **21**, 2632 (2009).
[344]S. Meister, S. B. Kim, J. J. Cha, H.-S. P. Wong, and Y. Cui, ACS Nano **5**, 2742 (2011).
[345]S. Chen, L. Jiang, M. Buckwell, X. Jing, Y. Ji, E. Grustan-Gutierrez, F. Hui, Y. Shi, M. Rommel, A. Paskaleva, G. Benstetter, W. H. Ng, A. Mehonic, A. J. Kenyon, and M. Lanza, Adv. Funct. Mater. **28**, 1802266 (2018).
[346]M. Buckwell, L. Montesi, S. Hudziak, A. Mehonic, and A. J. Kenyon, Nanoscale **7**, 18030 (2015).
[347]A. Mehonic, M. Buckwell, L. Montesi, M. S. Munde, D. Gao, S. Hudziak, R. J. Chater, S. Fearn, D. McPhail, M. Bosman, A. L. Shluger, and A. J. Kenyon, Adv. Mater. **28**, 7486 (2016).
[348]K. Szot, W. Speier, G. Bihlmayer, and R. Waser, Nat. Mater. **5**, 312 (2006).
[349]H. R. J. Cox, M. Buckwell, W. H. Ng, D. J. Mannion, A. Mehonic, P. R. Shearing, S. Fearn, and A. J. Kenyon, APL Mater. **9**, 111109 (2021).
[350]I. Valov and T. Tsuruoka, J. Phys. D: Appl. Phys. **51**, 413001 (2018).
[351]A. Mehonic, A. Vrajitoarea, S. Cueff, S. Hudziak, H. Howe, C. Labbé, R. Rizk, M. Pepper, and A. J. Kenyon, Sci. Rep. **3**, 2708 (2013).
[352]S. Stathopoulos, L. Michalas, A. Khiat, A. Serb, and T. Prodromakis, Sci. Rep. **9**, 19412 (2019).
[353]R. Yuan, Q. Duan, P. J. Tiw, G. Li, Z. Xiao, Z. Jing, K. Yang, C. Liu, C. Ge, R. Huang, and Y. Yang, Nat. Commun. **13**, 3973 (2022).
[354]J. del Valle, P. Salev, F. Tesler, N. M. Vargas, Y. Kalcheim, P. Wang, J. Trastoy, M.-H. Lee, G. Kassabian, J. G. Ramírez, M. J. Rozenberg, and I. K. Schuller, Nature **569**, 388 (2019).
[355]S. G. Sarwat, B. Kersting, T. Moraitis, V. P. Jonnalagadda, and A. Sebastian, Nat. Nanotechnol. **17**, 507 (2022).
[356]G. W. Burr, M. J. BrightSky, A. Sebastian, H.-Y. Cheng, J.-Y. Wu, S. Kim, N. E. Sosa, N. Papandreou, H.-L. Lung, H. Pozidis, E. Eleftheriou, and C. H. Lam, IEEE J. Emerging Sel. Top. Circuits Syst. **6**, 146 (2016).
[357]M. Salinga, B. Kersting, I. Ronneberger, V. P. Jonnalagadda, X. T. Vu, M. Le Gallo, I. Giannopoulos, O. Cojocaru-Mirédin, R. Mazzarello, and A. Sebastian, Nat. Mater. **17**, 681 (2018).
[358]F. Rao, K. Ding, Y. Zhou, Y. Zheng, M. Xia, S. Lv, Z. Song, S. Feng, I. Ronneberger, R. Mazzarello, W. Zhang, and E. Ma, Science **358**, 1423 (2017).
[359]S. Bhatti, R. Sbiaa, A. Hirohata, H. Ohno, S. Fukami, and S. N. Piramanayagam, Mater. Today **20**, 530 (2017).
[360]L. W. Martin and A. M. Rappe, Nat. Rev. Mater. **2**, 16087 (2016).
[361]J. Hwang, Y. Goh, and S. Jeon, Small **20**, 2305271 (2023).
[362]K.-H. Kim, I. Karpov, R. H. Olsson, and D. Jariwala, Nat. Nanotechnol. **18**, 422 (2023).
[363]A. Chanthbouala, V. Garcia, R. O. Cherifi, K. Bouzehouane, S. Fusil, X. Moya, S. Xavier, H. Yamada, C. Deranlot, N. D. Mathur, M. Bibes, A. Barthélémy, and J. Grollier, Nat. Mater. **11**, 860 (2012).
[364]M. Onen, N. Emond, B. Wang, D. Zhang, F. M. Ross, J. Li, B. Yildiz, and J. A. del Alamo, Science **377**, 539 (2022).
[365]M. L. Gallo, T. Tuma, F. Zipoli, A. Sebastian, and E. Eleftheriou, in *2016 46th European Solid-State Device Research Conference (ESSDERC)* (IEEE, 2016).
[366]A. Fukushima, T. Seki, K. Yakushiji, H. Kubota, H. Imamura, S. Yuasa, and K. Ando, Appl. Phys. Express **7**, 083001 (2014).
[367]D. Vodenicarevic, N. Locatelli, A. Mizrahi, J. S. Friedman, A. F. Vincent, M. Romera, A. Fukushima, K. Yakushiji, H. Kubota, S. Yuasa, S. Tiwari, J. Grollier, and D. Querlioz, Phys. Rev. Appl. **8**, 054045 (2017).
[368]W. Qian, X. Li, M. D. Riedel, K. Bazargan, and D. J. Lilja, IEEE Trans. Comput. **60**, 93 (2011).
[369]F. Neugebauer, I. Polian, and J. P. Hayes, in *2017 Euromicro Conference on Digital System Design (DSD)* (IEEE, 2017).
[370]K.-E. Harabi, T. Hirtzlin, C. Turck, E. Vianello, R. Laurent, J. Droulez, P. Bessière, J.-M. Portal, M. Bocquet, and D. Querlioz, Nat. Electron. **6**, 52 (2022).
[371]A. Sengupta, G. Srinivasan, D. Roy, and K. Roy, in *2018 IEEE International Electron Devices Meeting (IEDM)* (IEEE, 2018).
[372]K. Y. Camsari, R. Faria, B. M. Sutton, and S. Datta, Phys. Rev. X **7**, 031014 (2017).
[373]A. Mizrahi, T. Hirtzlin, A. Fukushima, H. Kubota, S. Yuasa, J. Grollier, and D. Querlioz, Nat. Commun. **9**, 1533 (2018).
[374]A. Sengupta, P. Panda, P. Wijesinghe, Y. Kim, and K. Roy, Sci. Rep. **6**, 30039 (2016).
[375]N. Mohseni, P. L. McMahon, and T. Byrnes, Nat. Rev. Phys. **4**, 363 (2022).
[376]A. Grimaldi, K. Selcuk, N. A. Aadit, K. Kobayashi, Q. Cao, S. Chowdhury, G. Finocchio, S. Kanai, H. Ohno, S. Fukami, and K. Y. Camsari, in *2022 International Electron Devices Meeting (IEDM)* (IEEE, 2022).
[377]N. A. Aadit, A. Grimaldi, M. Carpentieri, L. Theogarajan, J. M. Martinis, G. Finocchio, and K. Y. Camsari, Nat. Electron. **5**, 460 (2022).
[378]C.-Y. Huang, W. C. Shen, Y.-H. Tseng, Y.-C. King, and C.-J. Lin, IEEE Electron Device Lett. **33**, 1108 (2012).
[379]H. Nili, G. C. Adam, B. Hoskins, M. Prezioso, J. Kim, M. R. Mahmoodi, F. M. Bayat, O. Kavehei, and D. B. Strukov, Nat. Electron. **1**, 197 (2018).
[380]A. F. Vincent, J. Larroque, N. Locatelli, N. Ben Romdhane, O. Bichler, C. Gamrat, W. S. Zhao, J.-O. Klein, S. Galdin-Retailleau, and D. Querlioz, IEEE Trans. Biomed. Circuits Syst. **9**, 166 (2015).
[381]J. Z. Sun, Phys. Rev. B **62**, 570 (2000).
[382]J. Kaiser, W. A. Borders, K. Y. Camsari, S. Fukami, H. Ohno, and S. Datta, Phys. Rev. Appl. **17**, 014016 (2022).
[383]A. Sebastian, R. Pendurthi, A. Kozhakhmetov, N. Trainor, J. A. Robinson, J. M. Redwing, and S. Das, Nat. Commun. **13**, 6139 (2022).
[384]S. Liu, T. P. Xiao, J. Kwon, B. J. Debusschere, S. Agarwal, J. A. C. Incorvia, and C. H. Bennett, Front. Nanotechnol. **4**, 1021943 (2022).
[385]D. Bonnet, T. Hirtzlin, A. Majumdar, T. Dalgaty, E. Esmanhotto, V. Meli, N. Castellani, S. Martin, J.-F. Nodin, G. Bourgeois, J.-M. Portal, D. Querlioz, and E. Vianello, Nat. Commun. **14**, 7530 (2023).
[386]K. Roy, A. Jaiswal, and P. Panda, Nature **575**, 607 (2019).
[387]B. Wickramasinghe, S. S. Chowdhury, A. K. Kosta, W. Ponghiran, and K. Roy, IEEE Trans. Cognit. Dev. Syst. (published online) (2024).
[388]A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, in *Advances in Neural Information Processing*

Systems (Curran Associates, Inc., 2017), Vol. 30; arXiv:1706.03762v7 (2017), https://doi.org/10.48550/arXiv.1706.03762.
[389]T. Mikolov, M. Karafiát, L. Burget, J. Černocký, and S. Khudanpur, in *INTERSPEECH 2010* (ISCA, 2010).
[390]A. Graves, A.-r. Mohamed, and G. Hinton, in *2013 IEEE International Conference on Acoustics Speech and Signal Processing* (IEEE, 2013).
[391]S. Hochreiter and J. Schmidhuber, Neural Comput. **9**, 1735 (1997).
[392]K. Cho *et al.*, "On the properties of neural machine translation: Encoder–decoder approaches," in *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation* (Association for Computational Linguistics, Doha, Qatar, 2014), pp. 103–111; arXiv:1409.1259.
[393]J. Chung *et al.*, in NIPS 2014 Workshop on Deep Learning, 2014; arXiv:1412.3555.
[394]E. O. Neftci *et al.*, IEEE Signal Process. Mag. **36**(6), 51 (2019).
[395]S. S. Chowdhury, C. Lee, and K. Roy, Neurocomputing **464**, 83 (2021).
[396]W. Ponghiran and K. Roy, in *Proceedings of the AAAI Conference on Artificial Intelligence* (AAAI, 2022), pp. 8001–8008.
[397]A. Dosovitskiy *et al.*, in International Conference on Learning Representations, 2020; arXiv:2010.11929v2.
[398]P. U. Diehl, D. Neil, J. Binas, M. Cook, S.-C. Liu, and M. Pfeiffer, in *2015 International Joint Conference on Neural Networks (IJCNN)* (IEEE, 2015).
[399]C. Lee, S. S. Sarwar, P. Panda, G. Srinivasan, and K. Roy, Front. Neurosci. **14**, 119 (2020).
[400]J. Kim, H. Kim, S. Huh, J. Lee, and K. Choi, Neurocomputing **311**, 373 (2018).
[401]S. Park, S. Kim, H. Choe, and S. Yoon, "Fast and efficient information transmission with burst spikes in deep spiking neural networks," in *Proceedings of the 56th Annual Design Automation Conference* (ACM, Las Vegas NV USA, 2019), pp. 1–6.
[402]I. Garg, S. S. Chowdhury, and K. Roy, in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (IEEE, 2021).
[403]S. Park, S. Kim, B. Na, and S. Yoon, "T2FSNN: Deep spiking neural networks with time-to-first-spike coding," in *2020 57th ACM/IEEE Design Automation Conference (DAC)* (IEEE, San Francisco, CA, USA, 2020), pp. 1–6.
[404]B. Han and K. Roy, *Computer Vision ECCV 2020* (Springer International Publishing, 2020), pp. 388–404.
[405]N. Rathi and K. Roy, IEEE Trans. Neural Networks Learn. Syst. **34**, 3174 (2023).
[406]B. Rueckauer, I. A. Lungu, Y. Hu, M. Pfeiffer, and S. C. Liu, Front. Neurosci. **11**, 682 (2017).
[407]C. Brandli, R. Berner, M. Yang, S.-C. Liu, and T. Delbruck, IEEE J. Solid-State Circuits **49**, 2333 (2014).
[408]A. Zhu, L. Yuan, K. Chaney, and K. Daniilidis, in *Robotics: Science and Systems XIV* (Robotics: Science and Systems Foundation, 2018).
[409]A. Z. Zhu, L. Yuan, K. Chaney, and K. Daniilidis, in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (IEEE, 2019).
[410]A. K. Kosta and K. Roy, in *IEEE International Conference on Robotics and Automation (ICRA)* (IEEE, 2023), pp. 6021–6027; arXiv:2209.11741.
[411]A. Sengupta, Y. Ye, R. Wang, C. Liu, and K. Roy, Front. Neurosci. **13**, 95 (2019).
[412]J. Kaiser, H. Mostafa, and E. Neftci, Front. Neurosci. **14**, 424 (2020).
[413]D. Huh and T. J. Sejnowski, in Advances in Neural Information Processing Systems, 2018; arXiv:1706.04698.
[414]N. Rathi *et al.*, in International Conference on Learning Representations, 2019; arXiv:2005.01807.
[415]S. S. Chowdhury *et al.*, in *European Conference on Computer Vision* (Cham; Springer Nature, Switzerland, 2022), pp, 709–726; arXiv:2110.05929.
[416]G. Datta, Z. Liu, and P. A. Beerel, arXiv:2212.10170v1 [cs.CV] (2022).
[417]K. Suetake, S. I. Ikegawa, R. Saiin, and Y. Sawada, Neural Networks **159**, 208 (2023).
[418]P. O'Connor, E. Gavves, and M. Welling, "Training a spiking neural network with equilibrium propagation," in *The 22nd International Conference on Artificial Intelligence and Statistics* (PMLR, 2019), pp. 1516–1523.
[419]M. Xiao, Q. Meng, Z. Zhang, Y. Wang, and Z. Lin, "Training feedback spiking neural networks by implicit differentiation on the equilibrium state," in *Advances in Neural Information Processing Systems (NeurIPS 2021)* (Curran Associates, Inc., 2021), Vol. 34, pp. 14516–14528.
[420]N. Rathi, I. Chakraborty, A. Kosta, A. Sengupta, A. Ankit, P. Panda, and K. Roy, ACM Comput. Surv. **55**, 243 (2023).
[421]B. V. Benjamin, P. Gao, E. McQuinn, S. Choudhary, A. R. Chandrasekaran, J.-M. Bussat, R. Alvarez-Icaza, J. V. Arthur, P. A. Merolla, and K. Boahen, Proc. IEEE **102**, 699 (2014).
[422]D. G. Elliott, M. Stumm, W. M. Snelgrove, C. Cojocaru, and R. Mckenzie, IEEE Des. Test Comput. **16**, 32 (1999).
[423]A. Agrawal, M. Ali, M. Koo, N. Rathi, A. Jaiswal, and K. Roy, IEEE Solid-State Circuits Lett. **4**, 137 (2021).
[424]I. Chakraborty, A. Jaiswal, A. K. Saha, S. K. Gupta, and K. Roy, Appl. Phys. Rev. **7**, 021308 (2020).
[425]M. Hosomi, H. Yamagishi, T. Yamamoto, K. Bessho, Y. Higo, K. Yamane, H. Yamada, M. Shoji, H. Hachino, C. Fukumoto, H. Nagao, and H. Kano, in *IEEE International Electron Devices Meeting 2005. IEDM Technical Digest* (IEEE, 2005).
[426]A. Sengupta and K. Roy, Appl. Phys. Rev. **4**, 041105 (2017).
[427]T. Sharma, C. Wang, A. Agrawal, and K. Roy, in *2021 IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED)* (IEEE, 2021).
[428]P. Panda and K. Roy, "Unsupervised regenerative learning of hierarchical features in Spiking Deep Networks for object recognition," in *2016 International Joint Conference on Neural Networks (IJCNN)* (IEEE, Vancouver, BC, Canada, 2016), pp. 299–306; arXiv:1602.01510.
[429]Z. Sun and D. Ielmini, IEEE Trans. Circuits Syst. **69**, 3024 (2022).
[430]Z. Sun, S. Kvatinsky, X. Si, A. Mehonic, Y. Cai, and R. Huang, Nat. Electron. **6**, 823 (2023).
[431]M. Le Gallo, A. Sebastian, G. Cherubini, H. Giefers, and E. Eleftheriou, IEEE Trans. Electron Devices **65**, 4304 (2018).
[432]E. J. Fuller, S. T. Keene, A. Melianas, Z. Wang, S. Agarwal, Y. Li, Y. Tuchman, C. D. James, M. J. Marinella, J. J. Yang, A. Salleo, and A. A. Talin, Science **364**, 570 (2019).
[433]T. P. Xiao, B. Feinberg, C. H. Bennett, V. Agrawal, P. Saxena, V. Prabhakar, K. Ramkumar, H. Medu, V. Raghavan, R. Chettuvetty, S. Agarwal, and M. J. Marinella, IEEE Trans. Circuits Syst. **69**, 1480 (2022).
[434]L. Pan, P. Zuo, Y. Luo, Z. Sun, and R. Huang, arXiv:2401.10042v1 [cs.AR] (2024).
[435]Y. Luo, S. Wang, P. Zuo, Z. Sun, and R. Huang, IEEE Trans. Circuits Syst. **69**, 4367 (2022).
[436]Z. Sun, G. Pedretti, E. Ambrosi, A. Bricalli, W. Wang, and D. Ielmini, Proc. Natl. Acad. Sci. U. S. A. **116**, 4123 (2019).
[437]B. Feinberg, R. Wong, T. P. Xiao, C. H. Bennett, J. N. Rohan, E. G. Boman, M. J. Marinella, S. Agarwal, and E. Ipek, in *2021 IEEE International Symposium on High-Performance Computer Architecture (HPCA)* (IEEE, 2021).
[438]P. Mannocci, E. Melacarne, A. Pezzoli, G. Pedretti, C. Villa, F. Sancandi, U. Spagnolini, and D. Ielmini, in *2023 International Electron Devices Meeting (IEDM)* (IEEE, 2023).
[439]Q. Zeng, J. Liu, M. Jiang, J. Lan, Y. Gong, Z. Wang, Y. Li, C. Li, J. Ignowski, and K. Huang, IEEE Internet Things J. **11**, 5169 (2024).
[440]S. Wang, Y. Luo, P. Zuo, L. Pan, Y. Li, and Z. Sun, Sci. Adv. **9**, eadj2908 (2023).
[441]Z. Chen, Y. Zhou, H. Xu, Y. Fu, Y. Li, Y. He, X.-S. Miao, P. Mannocci, and D. Ielmini, in *2023 International Electron Devices Meeting (IEDM)* (IEEE, 2023).
[442]J. Woo and S. Yu, IEEE Nanotechnol. Mag. **12**, 36 (2018).
[443]S. Achour, R. Sarpeshkar, and M. C. Rinard, ACM SIGPLAN Not. **51**, 177 (2016).
[444]J. Zhao, S. Huang, O. Yousuf, Y. Gao, B. D. Hoskins, and G. C. Adam, Front. Neurosci. **15**, 749811 (2021).
[445]W. Yi, F. Perner, M. S. Qureshi, H. Abdalla, M. D. Pickett, J. J. Yang, M.-X. M. Zhang, G. Medeiros-Ribeiro, and R. S. Williams, Appl. Phys. A **102**, 973 (2011).
[446]W. A. Borders, A. Madhavan, M. W. Daniels, V. Georgiou, M. Lueker-Boden, T. S. Santos, P. M. Braganca, M. D. Stiles, J. J. McClelland, and B. D. Hoskins, arXiv:2312.06446v1 [cs.ET] (2023).
[447]B. Feinberg, U. K. R. Vengalam, N. Whitehair, S. Wang, and E. Ipek, in *2018 ACM/IEEE 45th Annual International Symposium on Computer Architecture (ISCA)* (IEEE, 2018).
[448]P. Zuo, Z. Sun, and R. Huang, IEEE Trans. Circuits Syst. **70**, 2335 (2023).

[449]P. Mannocci, E. Giannone, and D. Ielmini, in *2023 IEEE International Conference on Metrology for Extended Reality, Artificial Intelligence and Neural Engineering (MetroXRAINE)* (IEEE, 2023).
[450]P. Mannocci and D. Ielmini, IEEE J. Explor. Solid-State Comput. Devices Circuits **9**, 47 (2023).
[451]S. Shapero, A. S. Charles, C. J. Rozell, and P. Hasler, IEEE J. Emerging Sel. Top. Circuits Syst. **2**, 530 (2012).
[452]S. Rehman, M. F. Khan, H.-D. Kim, and S. Kim, Nat. Commun. **13**, 2804 (2022).
[453]C. E. Graves, C. Li, G. Pedretti, and J. P. Strachan, *Memristor Computing Systems* (Springer International Publishing, 2022), pp. 105–139.
[454]C. Li, C. E. Graves, X. Sheng, D. Miller, M. Foltin, G. Pedretti, and J. P. Strachan, Nat. Commun. **11**, 1638 (2020).
[455]X. S. Hu, M. Niemier, A. Kazemi, A. F. Laguna, K. Ni, R. Rajaei, M. M. Sharifi, and X. Yin, in *2021 IEEE International Electron Devices Meeting (IEDM)* (IEEE, 2021).
[456]L. Grinsztajn, E. Oyallon, and G. Varoquaux, in NeurIPS 2022 Datasets and Benchmarks Track, 2022.
[457]Z. Xie, W. Dong, J. Liu, H. Liu, and D. Li, in *Proceedings of the Sixteenth European Conference on Computer Systems* (ACM, 2021).
[458]G. Pedretti, C. E. Graves, S. Serebryakov, R. Mao, X. Sheng, M. Foltin, C. Li, and J. P. Strachan, Nat. Commun. **12**, 5806 (2021).
[459]A. Shafiee, A. Nag, N. Muralimanohar, R. Balasubramonian, J. P. Strachan, M. Hu, R. S. Williams, and V. Srikumar, in *2016 ACM/IEEE 43rd Annual International Symposium on Computer Architecture (ISCA)* (IEEE, 2016).
[460]G. Pedretti, J. Moon, P. Bruel, S. Serebryakov, R. M. Roth, L. Buonanno, A. Gajjar, T. Ziegler, C. Xu, M. Foltin, P. Faraboschi, J. Ignowski, and C. E. Graves, arXiv:2304.01285 (2023).
[461]K. Pagiamtzis and A. Sheikholeslami, IEEE J. Solid-State Circuits **41**, 712 (2006).
[462]G. Pedretti, C. E. Graves, T. Van Vaerenbergh, S. Serebryakov, M. Foltin, X. Sheng, R. Mao, C. Li, and J. P. Strachan, Adv. Electron. Mater. **8**, 2101198 (2022).
[463]H. Farzaneh, J. P. C. de Lima, M. Li, A. A. Khan, X. S. Hu, and J. Castrillon, "C4CAM: A compiler for CAM-based In-memory accelerators," in *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems* (ACM, La Jolla CA USA, 2024), Vol. 3, pp. 164–177.
[464]A. Bremler-Barr, D. Hay, D. Hendler, and R. M. Roth, IEEE/ACM Trans. Networking **18**, 1665 (2010).
[465]A. Lucas, Front. Phys. **2**, 5 (2014).
[466]J. J. Hopfield and D. W. Tank, Biol. Cybern. **52**, 141 (1985).
[467]E. Farhi, J. Goldstone, and S. Gutmann, arXiv:1411.4028v1 [quant-ph] (2014).
[468]M. Ercsey-Ravasz and Z. Toroczkai, Nat. Phys. **7**, 966 (2011).
[469]B. Molnár, F. Molnár, M. Varga, Z. Toroczkai, and M. Ercsey-Ravasz, Nat. Commun. **9**, 4864 (2018).
[470]T. J. Sejnowski, AIP Conf. Proc. **151**, 398 (1986).
[471]R. Hamerly, T. Inagaki, P. L. McMahon, D. Venturelli, A. Marandi, T. Onodera, E. Ng, C. Langrock, K. Inaba, T. Honjo, K. Enbutsu, T. Umeki, R. Kasahara, S. Utsunomiya, S. Kako, K. Kawarabayashi, R. L. Byer, M. M. Fejer, H. Mabuchi, D. Englund, E. Rieffel, H. Takesue, and Y. Yamamoto, "Experimental investigation of performance differences between coherent Ising machines and a quantum annealer," Sci. Adv. **5**(5), eaau0823 (2019).
[472]T.-J. Yang and V. Sze, in *2019 IEEE International Electron Devices Meeting (IEDM)* (IEEE, 2019), pp. 1–22.
[473]J. Deguchi, D. Miyashita, A. Maki, S. Sasaki, K. Nakata, and F. Tachibana, in *2019 IEEE International Electron Devices Meeting (IEDM)* (IEEE, 2019), pp. 4–22.
[474]H. Jia, M. Ozatay, Y. Tang, H. Valavi, R. Pathak, J. Lee, and N. Verma, IEEE J. Solid-State Circuits **57**, 198 (2022).
[475]L. Fick, S. Skrzyniarz, M. Parikh, M. B. Henry, and D. Fick, in *2022 IEEE International Solid-State Circuits Conference (ISSCC)* (IEEE, 2022).
[476]P. Narayanan, S. Ambrogio, A. Okazaki, K. Hosokawa, H. Tsai, A. Nomura, T. Yasuda, C. Mackin, S. C. Lewis, A. Friz, M. Ishii, Y. Kohda, H. Mori, K. Spoon, R. Khaddam-Aljameh, N. Saulnier, M. Bergendahl, J. Demarest, K. W. Brew, V. Chan, S. Choi, I. Ok, I. Ahsan, F. L. Lie, W. Haensch, V. Narayanan, and G. W. Burr, IEEE Trans. Electron Devices **68**, 6629 (2021).
[477]M. L. Gallo, R. Khaddam-Aljameh, M. Stanisavljevic, A. Vasilopoulos, B. Kersting, M. Dazzi, G. Karunaratne, M. Brändli, A. Singh, S. M. Müller, J. Büchel, X. Timoneda, V. Joshi, M. J. Rasch, U. Egger, A. Garofalo, A. Petropoulos, T. Antonakopoulos, K. Brew, S. Choi, I. Ok, T. Philip, V. Chan, C. Silvestre, I. Ahsan, N. Saulnier, V. Narayanan, P. A. Francese, E. Eleftheriou, and A. Sebastian, Nat. Electron. **6**, 680 (2023).
[478]S. Kim, M. Ishii, S. Lewis, T. Perri, M. BrightSky, W. Kim, R. Jordan, G. W. Burr, N. Sosa, A. Ray, J.-P. Han, C. Miller, K. Hosokawa, and C. Lam, in *2015 IEEE International Electron Devices Meeting (IEDM)* (IEEE, 2015).
[479]S. Kim, S. Kim, S. Um, S. Kim, K. Kim, and H.-J. Yoo, in *2022 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits)* (IEEE, 2022).
[480]M. Anderson, B. Chen, S. Chen, S. Deng, J. Fix, M. Gschwind, A. Kalaiah, C. Kim, J. Lee, J. Liang *et al.*, arXiv:2107.04140 (2021).
[481]R. Khaddam-Aljameh, M. Stanisavljevic, J. F. Mas, G. Karunaratne, M. Brandli, F. Liu, A. Singh, S. M. Muller, U. Egger, A. Petropoulos, T. Antonakopoulos, K. Brew, S. Choi, I. Ok, F. L. Lie, N. Saulnier, V. Chan, I. Ahsan, V. Narayanan, S. R. Nandakumar, M. L. Gallo, P. A. Francese, A. Sebastian, and E. Eleftheriou, IEEE J. Solid-State Circuits **57**, 1027 (2022).
[482]S. Jain, H. Tsai, C.-T. Chen, R. Muralidhar, I. Boybat, M. M. Frank, S. Wozniak, M. Stanisavljevic, P. Adusumilli, P. Narayanan, K. Hosokawa, M. Ishii, A. Kumar, V. Narayanan, and G. W. Burr, IEEE Trans. Very Large Scale Integr. Syst. **31**, 114 (2023).
[483]A. I. Khan, A. Daus, R. Islam, K. M. Neilson, H. R. Lee, H. S. P. Wong, and E. Pop, Science **373**, 1243 (2021).
[484]B. Prasad, S. Parkin, T. Prodromakis, C.-B. Eom, J. Sort, and J. L. MacManus-Driscoll, APL Mater. **10**, 090401 (2022).
[485]G. W. Burr, A. Sebastian, E. Vianello, R. Waser, and S. Parkin, APL Mater. **8**, 010401 (2020).
[486]T. Venkatesan and S. Williams, Appl. Phys. Rev. **9**, 010401 (2022).
[487]T. Chen, J. van Gelder, B. van de Ven, S. V. Amitonov, B. de Wilde, H.-C. Ruiz Euler, H. Broersma, P. A. Bobbert, F. A. Zwanenburg, and W. G. van der Wiel, Nature **577**, 341 (2020).
[488]R. A. John, Y. Demirağ, Y. Shynkarenko, Y. Berezovska, N. Ohannessian, M. Payvand, P. Zeng, M. I. Bodnarchuk, F. Krumeich, G. Kara, I. Shorubalko, M. V. Nair, G. A. Cooke, T. Lippert, G. Indiveri, and M. V. Kovalenko, Nat. Commun. **13**, 2074 (2022).
[489]N. Yantara, S. E. Ng, D. Sharma, B. Zhou, P. S. V. Sun, H. M. Chua, N. F. Jamaludin, A. Basu, and N. Mathews, Adv. Mater. **36**, 2305857 (2023).
[490]Y. van de Burgt, E. Lubberman, E. J. Fuller, S. T. Keene, G. C. Faria, S. Agarwal, M. J. Marinella, A. Alec Talin, and A. Salleo, Nat. Mater. **16**, 414 (2017).
[491]S. Goswami, R. Pramanick, A. Patra, S. P. Rath, M. Foltin, A. Ariando, D. Thompson, T. Venkatesan, S. Goswami, and R. S. Williams, Nature **597**, 51 (2021).
[492]S. Goswami, A. J. Matula, S. P. Rath, S. Hedström, S. Saha, M. Annamalai, D. Sengupta, A. Patra, S. Ghosh, H. Jani, S. Sarkar, M. R. Motapothula, C. A. Nijhuis, J. Martin, S. Goswami, V. S. Batista, and T. Venkatesan, Nat. Mater. **16**, 1216 (2017).
[493]M. Lanza, G. Molas, and I. Naveh, Nat. Electron. **6**, 260 (2023).
[494]E. J. Fuller, S. T. Keene, A. Melianas, Z. Wang, S. Agarwal, Y. Li, Y. Tuchman, C. D. James, M. J. Marinella, J. J. Yang, A. Salleo, and A. A. Talin, Science **364**, 570 (2019).

AFFILIATIONS

[1]Department of Electronic and Electrical Engineering, UCL, Torrington Place, London WC1E 7JE, United Kingdom
[2]Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Piazza L. da Vinci 32, 20133 Milano, Italy
[3]School of Electrical and Computer Engineering, Purdue University, West Lafayette, Indiana 47907, USA

[4]Department of Information Technology and Electrical Engineering, ETH Zurich, Gloriastrasse 35, 8092 Zurich, Switzerland
[5]Andrew and Erna Viterbi Faculty of Electrical and Computer Engineering, Technion–Israel Institute of Technology, Haifa 3200003, Israel
[6]Instituto de Microelectrónica de Sevilla (IMSE-CNM), CSIC and University of Seville, Avda. Américo Vespucio 28, 41092 Seville, Spain
[7]CNR-IMM Unit of Agrate Brianza, Via C. Olivetti 2, 20864 Agrate Brianza (MB), Italy
[8]Department of Physics, Loughborough University, Epinal Way, Loughborough LE11 3TU, United Kingdom
[9]STMicroelectronics, Knowledge Park III, Greater Noida 201304, India
[10]STMicroelectronics S.p.A., System Research and Application, Via Tolomeo 1, 20010 Cornaredo (MI), Italy
[11]Ming Hsieh Department of Electrical and Computer Engineering, University of Southern California, Los Angeles, California 90089, USA
[12]IBM Research Europe, Säumerstrasse 4, 8803 Rüschlikon, Switzerland
[13]NaMLab gGmbH, Nöthnitzer Stasse. 64a, 01187 Dresden, Germany
[14]Institute of Semiconductors and Microsystems, TU Dresden, 01062 Dresden, Germany
[15]Zernike Institute for Advanced Materials and CogniGron Center, University of Groningen, Nijenborgh 4, 9747 AG Groningen, The Netherlands
[16]CEA, CNRS, SPINTEC, Univ. Grenoble Alpes, rue des Martyrs, 38000 Grenoble, France
[17]Institute of Electronics, National Yang Ming Chiao Tung University, 1001 University Road, Hsinchu 30010, Taiwan
[18]Institute of Physics, University of Münster, Heisenbergstraße 11, 48149 Münster, Germany
[19]Physical Science and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Saudi Arabia
[20]Department of Materials Science and Engineering, National University of Singapore, Singapore 117575, Singapore
[21]Peter Grünberg Institute 7, Forschungszentrum Jülich GmbH, 52428 Jülich, Germany
[22]The Alan Turing Institute, British Library, 96 Euston Road, London NW1 2DB, United Kingdom
[23]School of Integrated Circuits, Peking University, Beijing 100871, China
[24]CNRS, Centre de Nanosciences et de Nanotechnologies, Université Paris-Saclay, 91120 Palaiseau, France
[25]CEA LETI, MINATEC Campus, 17 rue des Martyrs, Grenoble 38054, France
[26]Institute for Artificial Intelligence and School of Integrated Circuits, Peking University, No. 5 Yiheyuan Road, Haidian District, Beijing 100871, China
[27]Hewlett Packard Labs, 820 N McCartney Rd., Milpitas, California 95035, USA
[28]Peter Grünberg Institute (PGI-14) for Neuromorphic Compute Nodes, Forschungszentrum Jülich GmbH, 52428 Jülich, Germany
[29]Faculty of Electrical Engineering, RWTH Aachen University, 52062 Aachen, Germany
[30]Department of Electrical and Computer Engineering, University of California, Santa Barbara, California 93106, USA
[31]IBM Research-Almaden, 650 Harry Road San Jose, California 95120, USA
[32]Peter Gruenberg Institute 7 (PGI7) and JARA-FIT Research Centre, Juelich, Wilhelm-Johnen-Straße, 52425 Juelich, Germany
[33]Institute of Electrochemistry and Energy Systems, Bulgarian Academy of Sciences, "Acad. G. Bonchev" Str. No. 10, 1113 Sofia, Bulgaria
[34]Institute of Materials in Electrical Engineering and Information Technology 2 (IWE2), RWTH Aachen University, D-52056 Aachen, Germany

[a)]**Author to whom correspondence should be addressed:** adnan.mehonic.09@ucl.ac.uk