Memory-Centric Neuromorphic Computing With Nanodevices

Damien Querlioz, Tifenn Hirtzlin, Jacques-Olivier Klein, Etienne Nowak, Elisa Vianello, Marc Bocquet, Jean-Michel Portal, Miguel Romera, Philippe Talatchian, Julie Grollier

Memory-Centric Neuromorphic Computing With Nanodevices

Damien Querlioz*, Tifenn Hirtzlin*, Jacques-Olivier Klein*, Etienne Nowak†, Elisa Vianello†,
Marc Bocquet‡, Jean-Michel Portal‡, Miguel Romera§, Philippe Talatchian§, and Julie Grollier§
*Centre de Nanosciences et de Nanotechnologies, CNRS, Univ Paris-Sud, Université Paris-Saclay, 91120 Palaiseau, France
Email: damien.querlioz@u-psud.fr †CEA, LETI, Grenoble, France.
‡Institut Matériaux Microélectronique Nanosciences de Provence, Univ. Aix-Marseille et Toulon, CNRS, France
§UMP CNRS/Thales, Univ Paris-Sud, Université Paris-Saclay, 91120 Palaiseau, France, France.

***Abstract*—When performing artificial intelligence, CPUs and GPUs consume considerably more energy for moving data between logic and memory units than for doing arithmetic. Brains, by contrast, achieve superior energy efficiency by fusing logic and memory entirely. Currently, emerging memory nanodevices give us an opportunity to reproduce this concept. In this overview paper, we look at neuroscience inspiration to extract lessons on the design of memory-centric neuromorphic systems. We study the reliance of brains on approximate memory strategies, which can be reproduced for AI. We give the example of a hardware binarized neural network with resistive memory. Based on measurements on a hybrid CMOS/resistive memory chip, we see that such systems can exploit the properties of emerging memories without error correction, and achieve extremely high energy efficiency. Second, we see that brains use the physics of their memory devices in a way much richer than only storage. This can inspire radical electronic designs, where memory devices become a core part of computing. We have for example fabricated neural networks where magnetic memories are used as nonlinear oscillators to implement neurons, and their electrical couplings implement synapses. Such designs can harness the rich physics of nanodevices, without suffering from their drawbacks.**

I. Introduction

Through the developments of deep neural networks, artificial intelligence has made tremendous progress in recent years. Unfortunately, AI achievements come with a tremendous cost: energy consumption [1]. This energy cost limits AI use in embedded contexts, in many cases forcing them to rely on the cloud, and raises concerns about the growing power consumption of data centers. At the same time, the amount of intelligence than human brains achieve with a mean power consumption of twenty watts does not cease to amaze. This suggests the existence of fundamental differences between brains and computers when it comes to energy efficiency, and raises the hope that some low-power techniques of brains might be imitated.

One important difference between brains and computers is the way that they are dealing with memory. In computers, as well as graphics cards, computational units and memory are separated, both physically and conceptually. In recent years, the energy efficiency of logic has been increased much more efficiently than the one of memory access. We are now in a situation where memory access, whether on-chip or off-chip, use considerably more energy than arithmetic operations [2]. This situation is unfavorable to computations involving neural networks, as those rely on relatively simple arithmetic operations, but considerable amounts of parameters and variables, and therefore memory access: when computing neural networks, practically all the energy is consumed for moving data [3]. This is in sharp contrast with brains, where there is no memory array for storing parameters and variables such as synaptic weights and neuron values: all the memory is directly cointegrated with computation [4], [5], therefore avoiding the leasing source of energy consumption of computers on neural networks entirely.

Another strategy used by brains is the reliance on approximate computations. Neurons computes in an analog fashion based on nanodevices that are extremely noisy [6]. Computers and graphics cards normally compute using perfectly deterministic 32 or 64 bits floating-point arithmetics, which is considerably more precise than neurons, and this precision has an energy and resource cost. Exploiting the two brain-inspired ideas of bringing logic and memory closer and relying on approximate computation has led to a whole range of research for developing energy efficient AI hardware, in academia as well as industry. For example, the Tensor Processing Units developed by Google use on-chip memory in a way optimized for multiply-and-accumulate operations and rely on eight bits fixed point arithmetics [7], and can reduce considerably the energy consumption of neural network inference even on difficult tasks [8].

Unfortunately, approaches based on CMOS technology alone have a core limitation when it comes to bringing logic and memory closer: memory has to be implemented with static random access memory (SRAM), which is a large-area circuit and has not scaled efficiently in recent technology nodes [9]. By contrast, novel memory devices have been developed in recent years and are emerging as a solution. These devices, such as resistive oxide-based memory [10] or memristors [11], phase change memory [12] and spin torque magnetoresistive memory [13], provide fast, non-volatile and compact memory cells that can be embedded at the core of advanced CMOS processes. These technologies therefore appear ideal to implement brain-inspired AI hardware, and this topic has been subject to considerable research in recent years [12], [14], [15], with highly varying hypothesis.

In this overview paper, we present two approaches using nanodevices for memory-centric brain-inspired artificial intelligence. The first approach – implementing binarized neural networks with resistive memory (section II) – remains very true to the principles of digital electronics, but implementing the two brain-inspired ideas mentioned above. The second approach – implementing neurons with spin torque memories (section III) – goes a lot further in terms of bioinspiration and tries to use nanodevices more in the spirit of how brains are using them: memory nanodevices and their physics become a core part of the computation.

II. Implementing Binarized Neural Networks in Hardware

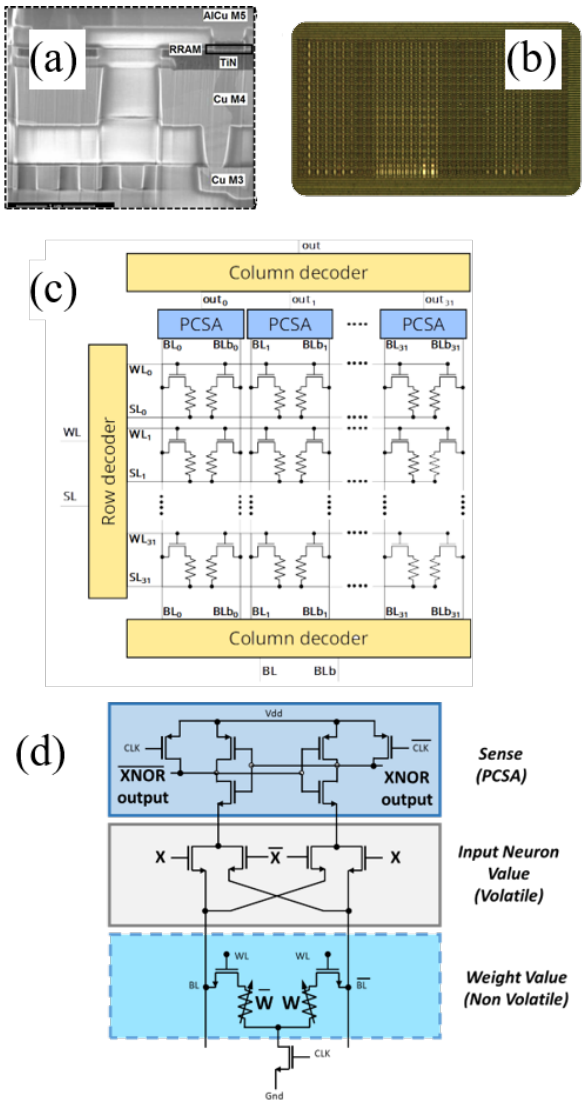

Fig. 1. (a) Electron microscopy image of a hafnium oxide RRAM cell in the CMOS backendof line. (b) Photograph and (c) simplified schematic of our test die. (d) Sense-amplifier (PCSA) circuit incorporating the XNOR operation of the BNN.

In this first project, we aim at implementing a neural network in hardware, with computation tightly integrated with memory, and all memory integrated on-chip. We propose using resistive memory for implementing synaptic weights, an idea widely investigated in the literature [4], [11], [12], [15]. However, a challenge is to be able to accommodate a whole model on chip. Fortunately, in recent years, considerable research has shown that at inference time, neural networks can function with low precision operations, which can considerably reduce the resource requirements for implementing a hardware neural network. The most extreme idea is to use Binarized Neural Networks, where both the synaptic weights and the neuronal activation values can assume only two values: $+1$ and -1 [16], [17]. Such neural networks require only one bit of memory per synapse and per neuron. Second, the multiplication between a neuronal value and a synaptic weights, normally the most expensive arithmetic operation in terms of area and energy is a reduced to logic operation between two binary values: an exclusive NOR (XNOR). Nevertheless, Binarized Neural Networks can achieve state-of-the-art performance on vision tasks [16]–[18]. For these reasons, these models are extremely attractive for inference hardware [15], [19]–[21]. It should be noted that during learning, to reach high accuracy, the synaptic weights should assume real values [16], [17]. The Binarized Neural Networks is therefore less attractive for learning-capable hardware.

Nevertheless, this approach comes with an important challenge: nanometer-scale memories are prove to device variability, which causes bit errors. We developed a specific hardware, presented in detail in [19], to investigate this issue and how to deal with it. Fig. 1(a) shows an electronic microscopy image of a hafnium-oxide based resistive memory device used in our work, positioned in the back-end-of-line of a 130 nanometer commercial CMOS process. Fig. 1(b) shows a photograph of the die with a one kilobit in-memory computing block, and Fig. 1(c) shows the simplified schematic of this block. In our design, each memory bit is stored using two memory devices, programmed in a complementary fashion: the combination of high and low resistance state means logic state one, and the inverse combination of low and high resistance means logic zero. The logic state is therefore obtained by comparing the resistance state of the two devices, using sense amplifiers integrated on-chip. This technique is more resilient to errors than the more traditional approaches where a single device is used by stored bit, and allows a reduction of the bit error rate due to device variation. We showed experimentally and theoretically that the benefits are similar to the use of single error correcting codes [19]. It should also be remarked, that binarized neural networks do not require the same determinism as conventional digital designs. Even if some bit errors remain, the accuracy of the overall system can be unaffected. For example, [22] shows that bit error rates as high as 10^{-3} have no impact on several vision tasks.

Another advantage of this approach is that it allows implementing ideas of logic-in-memory in a simple fashion. In Fig. 1(d), the XNOR operation, which implements multiplication, is directly performed within the sense amplifier, following an approach initially proposed in [23]. The XNOR

is performed at the same time as the read operation, and without area and energy overhead. The sum can then be implemented with low depth digital integer adders [21]. The resulting systems can lead to important savings in terms of energy and area with regards to non binarized ones [19], [21].

The fully digital approach of this work contrast with related works, but working in an analog fashion: resistive memories are programmed in an analog way, implementing real weights [11], [12] or binarized fashion [15], [24], and the neuronal values are calculated using Ohm's law and Kirchhoff's current law. With contrast to this approach, ours requires digital integer adders, but allows the use of fast and highly energy efficient sense amplifiers to read the state of the memories, while avoiding the need of any area and energy hungry analog circuit such as operational amplifier.

III. Computing with Physics: Using Spin Torque Nano-oscillators as Artificial Neurons

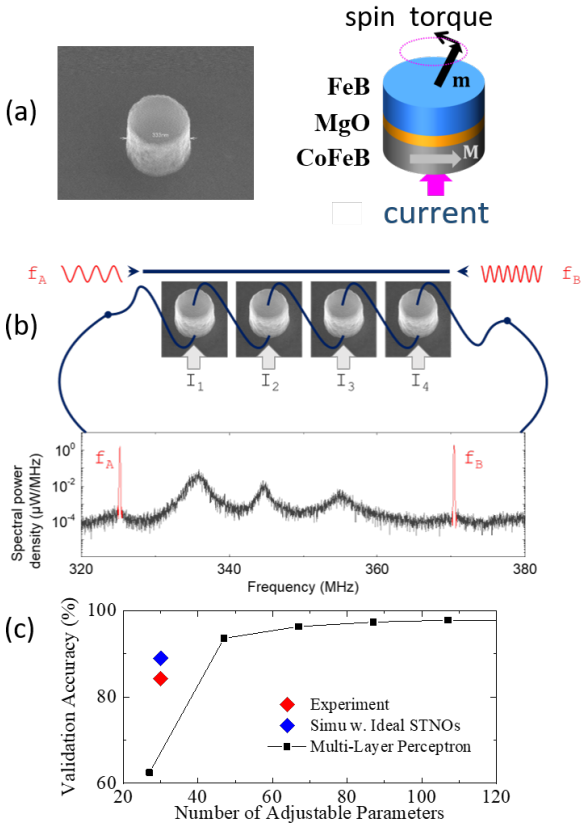

Fig. 2. (a) Electron microscopy and schematized visualization of a magnetic tunnel junction. (b) Schematic of our four-STNOs neural network, and power spectrum of a sample output of the network. (c) Validation accuracy of neural networks on a vowel recognition task: our experiment, simulation of our experiment with idealized STNOs, and standard multilayer perceptron with tanh activation function and softmax output.

In the brain, memory nanodevices are not used solely to store information. Synapses do store weights on the long term, but they also feature rich dynamics on multiple time scales, which are believed to be harnessed for processing information and far learning. Ion channels can be considered as memory elements that only function on millisecond timescales and are used as controlled current sources. Similarly, nanodevices envisioned to be used as memory in microelectronics feature highly complex physics and dynamical properties [14], [25], [26], and using them exclusively as storage elements as in the project reported in Sec. II can feel like a waste. Nevertheless, exploiting their dynamics for computing is an immense challenge due to the variability and electrical noise that is inherent in these devices due to their reliance on atomic-scale effects. As brains are very tolerant to approximate computing, neuromorphic computing seems the ideal venue for testing computing schemes that use the dynamical physics of nanodevices.

Here, we present a work where to choose to exacerbate the dynamical effects present in a memory device. We use spin torque nano-oscillators (STNOs, Fig. 2(a)). These devices use a magnetic tunnel junction, the same basic cell as spin-torque magnetoresistive memory (ST-MRAM), a vertical structures made of magnetic and non-magnetic layers, implementing two nanomagnets separated by a tunnel oxide. In a magnetic tunnel junction used for ST-MRAM, electrical currents can be used to switch the magnetization of one of the two nanomagnets between two stable states, the two memory states of the devices. Here, we use devices sized such as electrical currents do not switch this magnetization, but instead cause it to precess, giving rise to an oscillation of the electrical resistance of the device, and the generation of an alternating voltage (Fig. 2(a)). Such devices, which turn DC signals to AC, therefore behave as auto-oscillators, with a highly non-linear character, and a physics that is rich and very well understood. They can be tuned by varying electrical current and magnetic field, and several oscillators can also synchronize to each others as well as to external alternating signals (currents or magnetic fields) [27].

In neuroscience, neural networks have frequently been modeled as coupled non-linear auto-oscillators [28], and the computational power of such structures is recognized. Therefore, it is natural idea to try to use STNOs as artificial neurons in neuroscience-inspired neural networks. Our first work used a single STNO to implement a whole neural network [29]. The system used time-multiplexing, meaning that the STNO would emulate the neurons of the network sequentially in time. In this situation, synaptic coupling between the neurons emerges naturally from the sort term memory of the STNO. Such a neural network, implementing a concept inspired by reservoir computing, was able to recognize spoken digits at state-of-the-art performance, showing that the dynamics of nanodevices can be exploited despite device imperfections such as electrical noise [29].

This approach is however limited in the size and complexity of the neural network that can be implemented. Here, we focus on a more advanced work, first introduced in [30], where we use a network of electrically coupled STNOs, each implementing a neuron of a neural network (Fig. 2(b)). The electrical connection between the STNOs emulates synapses

connecting them. A strip line positioned on top of the STNOs allows to present the inputs as an alternating magnetic fields.

This simple system can emulate a full layer of a neural network, and in [30] was trained on a task of spoken vowel recognition. Fig. 2(c) reports the experimental recognition rate, compared with a simulation of the experiment where the STNOs are modeled as perfect oscillators, and standard software neural network employing $\tanh$ activation function and softmax outputs. We see that at equivalent number of parameters, the STNO neural network outperforms the traditional one. This is not surprising for the simulation that assumes perfect STNOs, as our neural network has more topological complexity and inherent neuronal non-linearity as the traditional one. What is interesting is that this benefit is not lost when we use real devices, which feature a high level of variability and phase noise. This highlights that the approach of benefiting from the dynamical physics of nanodevices, as in the brain, is feasible.

IV. Conclusion

This article summarizes two projects, where, inspired by the architecture of the brain, memory nanodevices have been put at the core of a computation scheme with the vision of achieving compact and energy efficient artificial intelligence. The first project – implementing binarized neural network with RRAM – remains true to the principles of digital VLSI, while incorporating the bioinspired ideas of logic and memory integration, low precision computation and intrinsic error tolerance. The second project – using spin oscillators as artifical neurons – goes a lot further by truly using memory devices as core computing elements, exploiting analog basic computation and the different physics of spin-electronic devices. It shows that it is possible to benefit of the complexity of nanodevice physics, without suffering from their drawbacks.

Future works should focus on scaling the approach to more application-ready levels. The first approach is of course closer to applications, as it benefits naturally from the achievements of VLSI, and it employs a form of neural networks that is also close to mainstream AI. Scaling the second approach comes with considerable more challenges, as both the technology and the associated neural network concepts are immature. It however allows dreaming about achieving brain-level integration and energy efficiency.

Acknowledgment

This work was supported by the European Research Council Grants NANOINFER (715872) and bioSPINspired (682955), and ANR grant NEURONIC (ANR-18-CE24-0009).

References

[1] Editorial, "Big data needs a hardware revolution," *Nature*, vol. 554, no. 7691, p. 145, Feb. 2018.

[2] A. Pedram, S. Richardson, M. Horowitz, S. Galal, and S. Kvatinsky, "Dark memory and accelerator-rich system optimization in the dark silicon era," *IEEE Design & Test*, vol. 34, no. 2, pp. 39–50, 2017.

[3] Y.-H. Chen, J. Emer, and V. Sze, "Using dataflow to optimize energy efficiency of deep neural network accelerators," *IEEE Micro*, vol. 37, no. 3, pp. 12–21, 2017.

[4] D. Querlioz *et al.*, "Bioinspired Programming of Memory Devices for Implementing an Inference Engine," *Proc. IEEE*, vol. 103, no. 8, pp. 1398–1416, 2015.

[5] G. Indiveri and S.-C. Liu, "Memory and information processing in neuromorphic systems," *Proc. IEEE*, vol. 103, no. 8, p. 1379, 2015.

[6] A. A. Faisal, L. P. Selen, and D. M. Wolpert, "Noise in the nervous system," *Nature reviews neuroscience*, vol. 9, no. 4, p. 292, 2008.

[7] N. P. Jouppi *et al.*, "In-datacenter performance analysis of a tensor processing unit," in *Proc. ISCA*. IEEE, 2017, pp. 1–12.

[8] D. Silver *et al.*, "Mastering the game of go without human knowledge," *Nature*, vol. 550, no. 7676, p. 354, 2017.

[9] T. Song *et al.*, "A 10 nm finfet 128 mb sram with assist adjustment system for power, performance, and area optimization," *IEEE Journal of Solid-State Circuits*, vol. 52, no. 1, pp. 240–249, 2016.

[10] D. R. B. Ly *et al.*, "Role of synaptic variability in resistive memory-based spiking neural networks with unsupervised learning," *J. Phys. D: Applied Physics*, 2018.

[11] M. Prezioso *et al.*, "Training and operation of an integrated neuromorphic network based on metal-oxide memristors," *Nature*, vol. 521, no. 7550, p. 61, 2015.

[12] S. Ambrogio *et al.*, "Equivalent-accuracy accelerated neural-network training using analogue memory," *Nature*, vol. 558, p. 60, 2018.

[13] O. Golonzka *et al.*, "Mram as embedded non-volatile memory solution for 22ffl finfet technology," in *IEDM Tech. Dig.*, 2018, pp. 18–1.

[14] J. Grollier *et al.*, "Spintronic nanodevices for bioinspired computing," *Proc. IEEE*, vol. 104, no. 10, p. 2024, 2016.

[15] S. Yu, "Neuro-inspired computing with emerging nonvolatile memorys," *Proc. IEEE*, vol. 106, no. 2, pp. 260–285, 2018.

[16] M. Courbariaux, I. Hubara, D. Soudry, R. El-Yaniv, and Y. Bengio, "Binarized neural networks: Training deep neural networks with weights and activations constrained to+ 1 or-1," *arXiv preprint arXiv:1602.02830*, 2016.

[17] M. Rastegari, V. Ordonez, J. Redmon, and A. Farhadi, "Xnor-net: Imagenet classification using binary convolutional neural networks," in *Proc. ECCV*. Springer, 2016, pp. 525–542.

[18] X. Lin, C. Zhao, and W. Pan, "Towards accurate binary convolutional neural network," in *Advances in Neural Information Processing Systems*, 2017, pp. 345–353.

[19] M. Bocquet, T. Hirztlin, J.-O. Klein, E. Nowak, E. Vianello, J.-M. Portal, and D. Querlioz, "In-memory and error-immune differential rram implementation of binarized deep neural networks," in *IEDM Tech. Dig.* IEEE, 2018, p. 20.6.1.

[20] E. Giacomin, T. Greenberg-Toledo, S. Kvatinsky, and P.-E. Gaillardon, "A robust digital rram-based convolutional block for low-power image processing and learning applications," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 66, no. 2, pp. 643–654, 2019.

[21] T. Hirtzlin, B. Penkovsky, M. Bocquet, J.-O. Klein, J.-M. Portal, and D. Querlioz, "Stochastic computing for hardware implementation of binarized neural networks," *arXiv preprint arXiv:1906.00915*, 2019.

[22] T. Hirtzlin, M. Bocquet, J.-O. Klein, E. Nowak, E. Vianello, J.-M. Portal, and D. Querlioz, "Outstanding bit error tolerance of resistive ram-based binarized neural networks," *arXiv preprint arXiv:1904.03652*, 2019.

[23] W. Zhao *et al.*, "Synchronous non-volatile logic gate design based on resistive switching memories," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 61, no. 2, pp. 443–454, 2014.

[24] S. Yu, Z. Li, P.-Y. Chen, H. Wu, B. Gao, D. Wang, W. Wu, and H. Qian, "Binary neural network with 16 mb rram macro chip for classification and online training," in *IEDM Tech. Dig.* IEEE, 2016, pp. 16–2.

[25] S. La Barbera *et al.*, "Interplay of multiple synaptic plasticity features in filamentary memristive devices for neuromorphic computing," *Scientific reports*, vol. 6, p. 39216, 2016.

[26] S. Kim, C. Du, P. Sheridan, W. Ma, S. Choi, and W. D. Lu, "Experimental demonstration of a second-order memristor and its ability to biorealistically implement synaptic plasticity," *Nano letters*, vol. 15, no. 3, pp. 2203–2211, 2015.

[27] N. Locatelli, V. Cros, and J. Grollier, "Spin-torque building blocks," *Nature materials*, vol. 13, no. 1, p. 11, 2014.

[28] A. Pikovsky, M. Rosenblum, J. Kurths, and J. Kurths, *Synchronization: a universal concept in nonlinear sciences*. Cambridge university press, 2003, vol. 12.

[29] J. Torrejon *et al.*, "Neuromorphic computing with nanoscale spintronic oscillators," *Nature*, vol. 547, no. 7664, p. 428, 2017.

[30] M. Romera *et al.*, "Vowel recognition with four coupled spin-torque nano-oscillators," *Nature*, vol. 563, no. 7730, p. 230, 2018.