Spintronic Devices as Key Elements for Energy-Efficient Neuroinspired Architectures

Nicolas Locatelli*, Adrien F. Vincent*, Alice Mizrahi†‡, Joseph S. Friedman*, Damir Vodencarevic*, Joo-Von Kim*, Jacques-Olivier Klein*, Weisheng Zhao*†, Julie Grollier†, and Damien Querlioz*

*Institut d’Electronique Fondamentale, Univ. Paris-Sud, CNRS, 91405 Orsay, France
Email: nicolas.locatelli@u-psud.fr, damien.querlioz@u-psud.fr
†Spintronics Interdisciplinary Center, Beihang University, Beijing, 100191, China
‡Unite Mixte de Physique CNRS/Thales and Universite Paris-Sud, 1 Ave. A. Fresnel, 91767 Palaiseau, France

Abstract—Processing the current deluge of data using conventional CMOS architectures requires a tremendous amount of energy, as it is inefficient for tasks such as data mining, recognition and synthesis. Alternative models of computation based on neuroinspiration can prove much more efficient for these kinds of tasks, but do not map ideally to traditional CMOS. Spintronics, by contrast, can bring features such as embedded nonvolatile memory and stochastic and memristive behavior, which, when associated with CMOS, can be key enablers for neuroinspired computing. In this paper, we explore different works that go in this direction. First, we illustrate how recent developments in embedded nonvolatile memory based on magnetic tunnel junctions (MTJs) can provide the large amount of nonvolatile memory required in neuro-inspired designs while avoiding Von Neumann bottleneck. Second, we show that recently developed spintronic memristors can implement artificial synapses for neuromorphic systems. With a more groundbreaking design, we show how the probabilistic writing of single MTJ bits can efficiently replace multi-level weighting for some classes of neuroinspired architectures. Finally, we show that a special class of MTJs can exhibit the phenomenon of stochastic resonance, a strategy used in biological systems to detect weak signals. These results suggest that the impact of spintronics extends beyond the traditional standalone and embedded memory markets.

I. INTRODUCTION

Emerging applications of electronics, such as Big Data and ubiquitous sensing, call for systems capable of handling large quantities of natural data with low power consumption. The relevant tasks, which rely on machine learning and can be qualified as “cognitive”, have been summarized as recognition, mining and synthesis [1]. These tasks do not map perfectly to traditional computers; most machine learning algorithms suffer heavily from the von Neumann bottleneck. In recent years, many researchers have advocated a bioinspired computing paradigm as a solution for implementing cognitive tasks with high energy efficiency [2]–[5]. However, CMOS does not provide all features necessary for ideal implementation of bioinspired architectures. In particular, fast and compact embedded nonvolatile memory and efficient random number generation are not available.

In parallel, spintronics (spin electronics) has emerged as a major advance in micro and nanoelectronics since the Nobel prize of Albert Fert and Peter Grünberg in 2007. Spintronics exploits the intrinsic magnetic moment (spin) of electrons in addition to their electrical charge to achieve novel features.

The prominent example is the magnetic tunnel junction, the flagship structure of spintronics, which exhibits a wide variety of complex and tunable behaviors [6], particularly a fast, compact and reliable binary nonvolatile memory. This memory function is envisioned for diverse applications as standalone and embedded memory, or in original logic-in-memory circuits [7]. Additionally, the features provided by spintronics share some features with biology [8]. Much recent work has investigated the use of nanotechnology for bioinspired computing [9]–[13], by mimicking biology’s use of nanoscale elements. In this paper, we show how spintronics fits specially well in this vision, and may complement CMOS ideally to provide the features necessary for emerging neuroinspired applications.

Here, after introducing several features that spintronics can bring to CMOS, we present ideas for using spintronics in bioinspired design. A form of hybrid logic differentiating between volatile and nonvolatile signals can implement cognitive algorithms. Spintronics devices may also be used as “synapses”. Finally, they may be used as stochastic oscillators that synchronize in a bioinspired way based on the phenomenon of “stochastic resonance”

II. SPINTRONIC BUILDING BLOCKS

Exploiting the spin degree of freedom of electrons permits numerous novel device features. Most spintronics devices are based on a magnetic tunnel junction structure (MTJ, Figure 1(a)). A MTJ consists in a patterned stack composed of layered magnetic and nonmagnetic materials. An MTJ consists of a patterned stack composed of layered magnetic
and nonmagnetic materials. It features a first nanomagnet with a pinned magnetization (fixed or pinned layer), and a second nanomagnet (free layer) with a magnetization that can be parallel (P) or antiparallel (AP) to the magnetization of the fixed layer. These two nanomagnets are separated by a thin tunnel barrier. The magnetization of the nanomagnets can be either in-plane, as illustrated in Figure 1, or perpendicular to the plane, in more recent designs.

Due to spin transport physics, the electrical resistance of the P and AP state differ (P is low resistance, AP is high resistance). This difference is quantified by the tunnel magnetoresistance ratio (TMR):

\[
TMR = \frac{R_{AP} - R_P}{R_P}.
\]

In typical devices, the TMR ranges between 30% and 150%. A magnetic field can be used to switch the state of the MTJ between the P and AP states. The first generation of magnetic random access memory (MRAM) was based on this principle; however, this leads to parasitic crosstalk between MTJs when the technology is scaled. A second spin transport effect, called “spin transfer torque”, allows the MTJ state to be switched using an electrical current flow through the MTJ. A positive current switches the MTJ from the AP to the P state, while a negative current switches the MTJ from the P to the AP state (Figure 1(b), Figure 2(a)). This is the basis of to the second generation of MRAM: Spin transfer torque MRAM (STT-MRAM) [14]. STT-MRAM provides compact nonvolatile memory with outstanding endurance and fast write speeds (in the range of 10ns) that can be embedded at the core of CMOS chips. Recent results suggest that writing speed could scales in the ns range [15], providing a relatively “universal” memory that is fast, reliable and nonvolatile.

A unique feature of MTJ-based memory is the stochastic nature of its switching (Figure 2(b)) [14], [16], [17]. This effect results from thermal effects in the MTJ switching process, which is well-studied and can be modeled accurately [18], [19]. When a short programming pulse is applied to a MTJ, there is only a probability of switching the MTJ. Therefore, we need to use long programming pulses to switch MTJs reliably. On the other hand, this stochastic switching make MTJs true nanoscale random number generator, as has been proved experimentally in Ref. [20].

While simple MTJs of the form of Figure 1 possess only two stable memory states, variations are currently developed to feature multibit information storage. This spintronic “memristor” behavior is produced by engineering the MTJ such that the free layer’s magnetization is divided into two domains: one domain is aligned to the fixed layer, the other domain is antiparallel to the fixed layer (Figure 2(c)). The domain position can be shifted using current pulses. This device has for example been proposed in Refs. [21], [22], and a demonstration has been realized experimentally in Ref. [23].

Another original application of MTJs is to allow switching in response to thermal noise, without applying a magnetic field or an electrical current (Figure 2(d)). These so-called super-paramagnetic junctions behave as stochastic oscillators with electrical resistance states analogous to a telegraph noise signal, with properties determined by the sizing of the MTJ. We will show that this feature of controlled telegraph noise can be useful for groundbreaking computing applications.

Finally, it should be mentioned that MTJs can be used in many additional systems not discussed in the present work: e.g. as a magnetic field detector, microwave oscillator and receiver, or spin wave emitter [6]. In summary, MTJs are simple compact device with a wide range of potential applications due to their rich physics. The rest of this paper discusses techniques for integrating spintronic devices with CMOS circuits to develop efficient neuroinspired architectures.

### III. Non-Volatile Logic Gates for Cognitive Computing

Magnetic tunnel junctions constitute a CMOS-compatible technology and can be embedded at the core of CMOS. This allows considering hybrid circuits that benefit from the advantages of both CMOS and spintronics, as its nonvolatility. An example of such approach is presented in [24] and Figure 3. Zhao et al. propose logic gates where some inputs are traditional volatile CMOS signals, and some inputs are stored in nonvolatile MTJs. Ref. [24] introduces a comprehensive methodology to design such gates. Figure 3 presents a full-adder designed using this methodology. The A and carry (Ci) inputs are volatile, the B input is nonvolatile, while outputs SUM and carry (C0) are volatile. The nonvolatile input B is stored in complementary MTJs, and can be changed by a standard MTJ write circuit. The outputs of the circuit are
read using a Precharge Sense Amplifiers (PCSA), which are fast and energy-efficient MTJ read circuit. The volatile inputs are inserted between the non volatile input and the PCSAs as part of a MOS tree that alters the MTJ read operation and implements the actual adder logic function.

Interestingly, once laid out and even including all the write circuitry for input B, the circuit is slightly more compact than a conventional CMOS full adder (18μm² vs. 20μm² in a 40nm technology). When input B is not changed, the circuit has nearly same delay (90ps vs. 75ps) and power consumption (2.0μW vs. 2.2μW) as a conventional CMOS full adder [24]. However, changing input B requires more time and power, as it consists in programming a MTJ.

This class of circuit is therefore useful when some inputs are fundamentally volatile, and some inputs are parameters that are rarely changed and need to be nonvolatile. Making such a separation in a general purpose computer is a difficult task. However, for a machine learning algorithm, this separation is intuitive. Machine learning algorithms largely differentiate between model parameters and actual inputs to be processed. For example, in a deep network, neuron states need to be updated frequently, while synaptic weights are not changed once the network has been trained. Putting the parameters as nonvolatile inputs could lead to a machine learning specialized circuit that avoids issues related to von Neumann bottleneck and is instant on/off.

IV. SPINTRONIC DEVICES AS ARTIFICIAL SYNAPSES

A. Multilevel Synapses

In recent years, many research groups have focused on using memristive devices as synapses. Synapses are the connections in the brain, and also feature a form of long term memory: a synapse possesses an analog synaptic weight that modulates the transmission between neurons. Synapses are a fundamental element in neuroinspired systems; however, CMOS implementations of learning-capable synapses are volatile, and are not compact [25], [26]. Memristive devices are nonvolatile resistive memories that usually features multilevel memory, and possess the capability to naturally emulate some properties of synapses, such as the capability of learning with “spike timing dependent plasticity” [9]–[13], [27]. They might therefore implement ideally the synapses of neuromorphic systems.

The spintronic memristor introduced in section 2 can naturally implement this vision, and this idea is currently widely investigated [21]–[23]. A feature of spintronic memristor with regards to other memristive devices is that the memory states are well defined by domain wall pinning, while they are analog (and harder to control) in other technologies. Spintronic memristors, with programming times in the range of nanoseconds, also constitute a fast technology. Additionally, the separation between low and high resistance states is typically smaller for spintronic memristor than for other technologies. This might require the use of adapted differential circuits for read operation.

B. Stochastic Synapses

As mentioned in section 2, a distinctive feature of MTJs is stochastic switching [14], [16], [18]. This usually constitutes a nightmare for circuit engineers, which calls for long programming pulses or specialized circuits solutions such as self-enabled programming [28]. However, when using MTJs as synapses this can be turned into a useful feature: stochastic programming of a memory device can be a particularly efficient way to implement stochastic learning rules, which have proven useful in some situations [29], [30]. Synapses programmed using stochastic learning rules need not necessarily be multilevel, and we can therefore use standard binary MTJs as synapses using this concept.

Figure 4 illustrates this breakthrough idea [31], [32]. Figure 4(a) presents experimental measurements of stochastic switching: we can see that the probability for a MTJ to switch when a programming pulse is applied can be adjusted by...
choosing pulse duration and voltage amplitude. This effect is now fully understood and can be modeled using analytical equation [19].

Figure 4(b) presents a simplified learning architecture that implements a stochastic spike timing dependent plasticity rule, described in more detail in [32]. The MTJs, organized as a crossbar, connect CMOS input neurons to CMOS output neurons, in a manner analogous to “feed-forward” neural networks. Such a system can be trained to perform tasks in an unsupervised manner. The CMOS input neurons present the inputs as asynchronous voltage spikes, which are transmitted to the output differently depending on the MTJ states, which therefore act as binary synaptic weight. The CMOS output neurons act as leaky integrate-and-fire neurons [5]. When an output neuron spikes, it applies a simple voltage waveform on the crossbar, as illustrated in Figure 4(b), which leads to stochastic programming of the MTJs and implements the stochastic simplified spike timing dependent plasticity learning rule.

This seemingly simple process allows the system to learn complex tasks. For example, if we present as input a video of vehicles passing on a freeway acquired from a bioinspired retina [33], each output neuron naturally specializes to a particular lane of the freeway and becomes a vehicle detector for this lane. This is illustrated in Figure 4(c). Each subimage is a two-dimensional representation of the states of the MTJs connected to one output neuron; it is apparent that each output neuron is sensitive to one lane of the freeway. More neurons are sensitive to the four inward lanes than the two outward lanes, because fewer vehicles are driving on the two outward lanes. The system can thus be used as a vehicle counter, with detection rate on the four inward lanes higher than 95%, and a number of false detection smaller than 5%.

Interestingly, such a system is extremely robust to device variation, although these variations make the switching probabilities of the devices programmed in the same conditions extremely variable [31], [32]. The robustness to device variation, a common issue in nanoscale devices, makes a strong argument for using bioinspiration to better exploit emerging spintronics devices.

V. SPINTRONIC DEVICES FOR STOCHASTIC RESONANCE

The sensitivity of the spintronic nano-devices to thermal fluctuations can also be useful to harvest thermal noise energy in electronic systems. As already mentioned in section 2, when, by proper design, reducing the energy barrier separating the two P and AP states to a few times the thermal energy, a magnetic tunnel junction undergoes random switching between those states due to thermal fluctuations, therefore behaving as a stochastic oscillator [36], [37]. The MTJ’s free layer is then said to be superparamagnetic.

Such stochastic dynamics can lead to complex behaviors, relevant to bioinspired applications. Recent works have demonstrated the capacity of superparamagnetic tunnel junctions to naturally achieve stochastic resonance [34], [35], [38]. By properly setting the barrier height in respect to the thermal noise amplitude, the system exhibits an increased sensitivity to sub-threshold inputs mediated by the noise. For small amplitude periodic excitations, the mean frequency of the system response shows a strong correlation with the input frequency, demonstrating noise-enhanced synchronization.

In Figure 5, a square current signal with amplitude one sixth of the MTJ critical current and varying frequency is applied to a $60 \times 180 \text{ nm}^2$ superparamagnetic tunnel junction [34]. When the period of the input signal is small compared to the mean switching time $\tau$ of the MTJ (high frequency),
there is a low probability that a switch of the MTJ occurs during a half-period (Figure 5(b)). However, when the input frequency is decreased and the half-period compares to a few times $\tau$, the probability that an input oscillation triggers a single back-and-forth switch of the MTJ becomes increasingly significant (Figure 5(c)). When the frequency is decreased further, supplementary short glitches are then likely to occur due to the high instability of the device (Figure 5(d)).

As the resulting behavior, under a critical frequency, the switching frequency of the MTJ starts to correlate to the excitation frequency (Figure 5(e)) and the matching time between the input and output gradually increases (Figure 5(f)). Last but not least, this critical frequency increases as the device stability decreases, and these results are consistent with general stochastic resonance theory [35].

Stochastic resonance and its variations have been suggested as a strategy used by the brain to achieve low-power sensing in noisy environments [39], [40], and has also been proposed for applications [41], [42]. Nevertheless, harnessing the stochastic resonance property of superparamagnetic MTJs for applications will require further research.

VI. CONCLUSION

In this work, we have introduced the features that spin electronics can bring to CMOS and how they can enhance functionality of neuroinspired circuits. Several ideas have been introduced, exploiting variations of the flagship device of spin electronics: the MTJ. Logic gates with hybrid volatile and nonvolatile behaviors are especially appropriate to implement cognitive systems. Spintronic memristors and binary stochastic magnetic tunnel junctions may be used as synapses for systems capable of learning. Superparamagnetic MTJs exhibit stochastic resonance that could be used for sensing systems.

There are fundamental reasons for which the MTJ in its different flavors fits especially well in a bioinspired design. As a compact nanodevice with rich physics, switching characteristics enhanced by thermal noise and nonvolatility, it is more reminiscent of the nanodevices used by biology (ion channels, synapses) than traditional electron devices. All these results and considerations suggest that the impact of MTJs on micro and nanoelectronics may go further than traditional embedded and standalone applications, although considerable work is still needed to develop the alternative paradigms of computation of the future involving spintronics.

ACKNOWLEDGMENTS

This work was supported by the ANR COGNISPIN (ANR-13-JS03-0004-01) and the FP7 ICT BAMBI (FP7-ICT-2013-C) projects, the CNRS/MI DEFI NANO program and the European Research Council (ERC NanoBrain 2010 Stg 259068). The authors would like to thank T. Devolder, D. Ravelosona, P. Bessiere and J. Drozlleiz for fruitful discussions.

REFERENCES


