Nanocomposants émergents pour l'ingénierie neuromorphique

Fabien Alibart

UNIVERSITE DE LILLE – Ecole doctorale ENGSYS – Institut d'électronique, de microélectronique et de Nanotechnologies

Nanocomposants émergents pour l'ingénierie neuromorphique

Mémoire présenté pour obtenir l'Habilitation à Diriger des Recherches

Spécialité: Sciences de l'information et des systemes

Fabien Alibart

21/10/2022

Composition du Jury :

Pr. Kamal LMIMOUNI, Professeur des Universités, IEMN-Université de Lille	*Président*
Dr. Julie GROLLIER, Directrice de Recherche, CNRS-Thales	*Rapporteure*
Dr. Marie-Paule BESLAND, Directrice de Recherche, IMN-CNRS	*Rapporteure*
Pr. Sylvain SAIGHI, Professeur des Universités, IMS-Université de Bordeaux	*Rapporteur*
Dr. Yannick COFFINIER, Directeur de Recherche, IEMN-CNRS	*Examinateur*
Dr. Damien QUERLIOZ, Chargé de Recherche, C2N-CNRS	*Examinateur*
Dr. Dominique VUILLAUME, Directeur de Recherche, IEMN-CNRS	*Garant*

Table des matières

Note introductive du document :

Ce manuscrit propose de présenter un résumé de mes activités de recherche couvrant la période de 2012 à 2022. Ces travaux ont été réalisés au CNRS à l'Institut d'Électronique, de Microélectronique et de Nanotechnologies depuis 2012 et au Laboratoire Nanotechnologies et Nanosystèmes de 2017 à 2020. A travers ce manuscrit, je propose d'illustrer mes travaux centrés sur l'électronique neuromorphique et ayant cherchés à étudier comment différentes technologies pouvaient être utilisées dans ce contexte. Notamment, les composants de type memristor à base de TiO2 m'ont permis une approche classique en microélectronique visant à travailler sur la montée en maturité d'un composant et son intégration au niveau circuit et systèmes pour la réalisation de puces neuromorphiques. Un deuxième volet de mes travaux s'est intéressé à utiliser des mécanismes non-conventionnels observés dans les nanotechnologies pour réaliser des fonctions innovantes en électronique neuromorphique. Ces travaux sont restés à un niveau très amont visant principalement les preuves de principe et se sont attachés à diversifier les matériaux et composants, depuis les conducteurs ioniques aux matériaux organiques et des composants mémoires aux transistors électrochimiques. Enfin, le dernier volet de mes travaux concerne une approche largement interdisciplinaire combinant plusieurs thématiques autour de l'électronique neuromorphique. Je m'intéresse à l'utilisation de concepts de traitement de l'information issus de l'électronique neuromorphique (ou bio-inspirée) pour le fonctionnement de réseaux de capteurs organiques couplés aux réseaux de neurones biologiques. Ces travaux ont été initiés depuis 2015 dans le cadre d'une collaboration avec le laboratoire JPArc (LilleNeuroCog) et sont au cœur de mon projet de recherche actuel et futur.

Le premier chapitre se présente comme un article de perspective sur l'électronique neuromorphique. Il présente un état de l'art superficiel du domaine et s'attache à identifier les grands enjeux et objectifs sous différents angles. Ce chapitre propose une analyse suivant trois axes principaux : (i) une comparaison aux réseaux de neurones artificiels, (ii) les enjeux de l'implémentation matérielle et (iii) les perspectives offertes par la biologie.

Le deuxième chapitre est un article de revue publié en 2020 qui propose une synthèse des différents enjeux liés a l'intégration des memristors sur CMOS. Au niveau applicatif, les memristors sont ici considérés pour des applications de type réseaux de neurones statiques mais les défis identifiés pour les systèmes hybrides CMOS/memristors restent valides pour l'électronique neuromorphique. Ce chapitre propose une structure classique depuis les composants, l'intégration au niveau circuit, jusqu'aux défis au niveau système. Il permet notamment de montrer que l'innovation attendue pour les systèmes de type calcul en mémoire à partir de composants memristors nécessite un travail largement interdisciplinaire allant du matériau au système de traitement de l'information. Ce travail a été réalisé principalement en collaboration avec Amirali Amirsoleimani.

Le chapitre 3 a été publié en 2015 et couvre les travaux de thèse de Selina La Barbera. L'idée maitresse de ces travaux était d'exploiter la physique des composants mémoires filamentaires pour réaliser différentes formes de plasticité synaptique. Notamment, utiliser la volatilité de ces cellules mémoires liée à l'instabilité du filament conducteur a permis de mimer les mécanismes de plasticité court terme (STP) et long terme (LTP). Ces travaux ont été poursuivis ensuite pour étudier comment ces mécanismes pouvaient être utilisés pour réaliser des apprentissages non-supervisés.

Le chapitre 4 est un article publié en 2018 montrant comment les concepts issus de l'électronique neuromorphique peuvent être utilisés dans le contexte d'un réseau de capteurs ioniques de type transistors électrochimiques. Ces travaux montrent comment la dynamique des capteurs et leur variabilité permettent d'implémenter des fonctions non-triviales de classification de signaux dynamiques. Ces travaux ont été réalisés en collaboration principalement avec Sebastien Pecqueur.

Enfin, le chapitre 5 présente mon projet de recherche qui s'intéresse à utiliser le traitement de l'information neuromorphique et bio-inspiré pour développer des interfaces innovantes aux réseaux de neurones biologiques. L'idée directrice de ces travaux et d'utiliser un paradigme de traitement de l'information le plus proche des systèmes biologiques et d'intégrer des fonctions de traitement des signaux directement au niveau de l'interface des réseaux de neurones biologiques.

1.CHAPTER 1

Neuromorphic computing and engineering overview

1.1. INTRODUCTION

What is really intelligence? This simple question is today stimulating multiple answers, which depend strongly on the angle used to analyze it. From a human-centered approach, intelligence is associated to the ability of human to formulate complex ideas, understand non-trivial mechanisms or planned elaborated strategies to anticipate future events. From a biological viewpoint, intelligence can appear through multiple forms such as collective behaviors in animals and vegetals, or ability to optimized resources for survival of all living species. But todays, this question is not restricted anymore to living organisms and we start to believe that intelligence could be embedded in artificial objects. The Artificial Intelligence revolution (AI), started decades ago with the first computers, offers a new substrate for researchers to tackle this question: in addition to the philosophical or biological angles, engineering is now a new domain in which intelligence could be considered. Indeed, these past years have seen impressive progresses in this direction with computers solving complex problems such as the Go game played by AlphaGo, image recognition surpassing human capacity in the ImageNet challenge or autonomous vehicles and robots evolving in real-life environment. However, this story is not only an engineering question. It is the result of the integration of multiple discoveries, from biological and computational neurosciences, computer sciences and mathematics, to material and physical sciences. Consequently, any attempt to expose how engineering is today progressing toward the development of intelligent systems should consider an interdisciplinary approach and could not be limited to the development of hardware materials.

In this chapter, we will present an overview of a specific domain that is intimately linked to AI. Neuromorphic computing and engineering (NCE), a term coined by Carver Mead in the 70s, is indeed emerging as a central aspect of AI and is bridging together multiple scientific domains. Progresses in AI have been deeply marked by machine learning in general, and Artificial Neural Networks (ANNs) in particular. ANNs, and their more recent development toward deep networks were at the origin of the second AI revolution that occurs in the 2010's. Inspired by the computing principles of the brain, ANNs rely on neurons (activation functions) and synapses (weighted connections) to compute data. They are also integrating the key ingredient of learning through mathematical models to define the synaptic weights. Neuromorphic computing is based on the same key ingredients but is using some very distinctive elements. Section 2 will present how NCE is capitalizing on spike encoding to represent information and what is the impact of this choice on data processing. Notably, spike encoding implies deep modification on the activation function of neurons, which represents a significant difference in between Spiking Neural Networks (SNNs) and ANNs. We will present in section 3 the incidences of spike coding on learning rules used in SNNs. In particular, we will expose the main learning rules used in SNNs and what are the most promising strategies. Another important distinction of NCE with machine learning is its profound connection to the hardware used to compute information. This aspect was already present in the pioneering work of Carver Mead, which considered the analogy between an ionic channel and a transistor as a foundation for NCE. This leads today to strongly hardware-oriented strategies for NCE that will be presented through two main point of view. Section 4 will put the emphasis on the energy consumption challenge that NCE is trying to solve. This energy consumption is directly linked to the hardware substrates used to implement SNNs. Section 5 will present

the challenges of integration to reach the density of components (i.e. neurons and synapses) comparable to what biology can do.

1.2. DATA REPRESENTATION IN SNNS.

One of the major difference that we can recognize today in between ANNs and SNNs is based on the data encoding used in both approaches. In the one hand, ANNs are computing data with analog values through their layered structures. Input data can be considered as vectors $\{x_i\} \epsilon R^n$ presented at the input layer. The synaptic conductances in between two layers corresponds to matrices W = (w_{ij}). The vector.matrix operation is the essential operation realized to compute the output vector that is passed through the neurons activation function to generate the input vector of the next layer. In this sense, ANNs can be considered as frame-based operator where data are organized into series of vectors. Consequently, ANNs are clocked systems where the notion of throughput (number of operations per second) is a relevant metric. In the other hand, SNNs are encoding continuous analog information through spikes. The simplest data representation corresponds to a rate-coding scheme, i.e. the analog value of the signal carrying information (or strength of a stimuli) is associated to the average frequency of the train of pulses. The neuron can then transmit some analog signals through its mean firing rate (figure 1a). A second coding scheme is known as temporal-coding in which each individual neuron is using the timing of the spike with respect to others neurons in order to encode the signal. The first spike is carrying a strong analog value while the later ones are associated to smaller analog values (figure 1b). Both coding schemes are used to encode continuous signals, at the opposite to "frame-based" ANNs. It is to note that spike coding raises an important issues for data representation. Both rate-based and time-based are somehow ubiquitous in real spiking signals, but moving from one representation to the other is not straightforward. It turns out that an alternative description of the spike encoding that could describe both strategies would be to consider a probability of spikes. Quantum physics theory, where both the time and localization of a particle cannot be known at the same time, but only its probability of presence, is an interesting analogy.

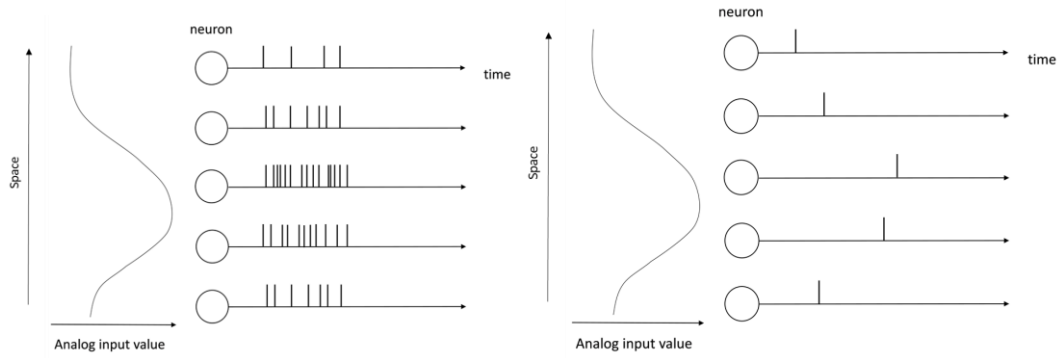

Figure 1: (a) Rate-based representation of analog signals. (b) Time-based representation of analog signals

In SNNs, spikes are discrete events, which could be associated to a digital value (i.e. "0" or "1"). Their temporal organization is carrying the analog features of the signal. In order to decode this analog information, neurons rely on the integration of the spikes. Various neurons models are used for NCE, with Leaky Integrate and Fire (LIF) being the most popular [2]. Each incoming spike contributes to increase the membrane potential of the neuron (V_{mem}). V_{mem} tends to leak with a given time constant. When the potential V_{mem} exceeds a threshold, the neuron fires an output spike and the membrane potential is reset to its resting state. This mechanism means that the neuron in itself is responsible for holding in its membrane potential part of the analog information (at least until V_{mem} leaks completely or a spike is emitted). This mechanism is often associated to a memory effect of the neuron. To some extends, an additionnal memory effect, which we describe here as the trace of the information in the network, can be identified through the recurrent connections and the delay of propagation. When propagating in the network and finally coming back to the

same place, a spike can "live" in a local point (i.e. a specific neuron) for a given duration. These considerations are used here to highlight a major difference in between ANNs and SNNs operation, which is intimately linked to the data representation. ANNs are data-driven system where static neurons and synapses compute the analog signals, only data are modified (this holds for inference only, not learning obviously). At the opposite, in SNNs, data are represented *in* the network through the combination of various dynamics of the elements that compose the SNNs. This representation of data (i.e. data are *in* the computing elements) is a major departure from the traditional way of considering data processing.

The spike encoding is an essential aspect that is limiting the development of SNNs. The conservative way of dealing with data processing is to consider data from sensors separately from the computing elements. However, as pointed out by H. Barlow 60 years ago [3]: *"it is foolish to investigate sensory mechanisms blindly—one must also look at the ways in which animals make use of their senses".* In other words, the way data are encoded at the sensory level cannot be separated from the higher computing levels. Along this line, developing SNNs for processing requires converting analog sensory signals into spikes. This analog-to-spike conversion is often under considered but is in fact a critical element of the processing itself. During the spike conversion, either the totality of the signal is converted or features from the signal can be enhanced or suppressed in order to convey only the meaningful part of information. For example, the simplest data conversion is to convert an analog signal into a train of spikes following the rate-coding strategy. In this technic, the limitation is mostly on the sampling of the analog signal since a minimal interval is required to define a mean frequency, but all information is transmitted without discrimination. More elaborated spike conversion are considering a more event-driven conversion by associating a spike to an *a-priori* important aspect of the signal such as a change in intensity (see for example BSA technic [4]). The latest progresses in this direction are now trying to develop finer features extraction such as in sparse coding strategies where filters of features need to be learn from the signal based on relevant criterion such as performances and sparsity.
Ultimately, NCE needs to consider not only SNN for processing but also the sensory elements of the circuit [5]. In this direction, the most significant achievements have been done with event-based camera that are emulating vision. The same bio-inspired approach has been applied to artificial cochlea implementation for sound processing. There is still lots of room for innovation in this domain if we think about the different sensory modalities used by living species to interact with their environment or to the large amount of sensors that are currently deployed on various Edge Computing applications.

Nevertheless, NCE is maybe somehow overestimating the importance of spikes. Analysis of electrophysiological signals reveals that signals in the brain present complex components such as subthreshold membrane potentials changes, collective Local Field Potential (LFP) resulting from the synchrony of populations of cells, or low frequency oscillations that are carrying an important part of the information (at least for some specific tasks). These different data representation are not integrated in current SNNs that are considering the spikes as the quantum of information. More recent progresses in biological neurosciences are now putting the emphasis on the role of additional elements to neurons to process / transmit information. The example of the tri-partite synapse where the astrocyte supervised the synaptic activity is a striking example of this issue. This could be extended further to the role of multiple chemical such as hormones in the way information is represented and processed in biological networks. The question that needs to be answered is: "is the spike enough to reproduce the complexity and performances of biological brains?"

1.3. LEARNING RULES IN SNNS

One of the keys of success of ANNs, and deep network in particular, is the backpropagation of error algorithm (backprop). Nevertheless, backprop required to be able to differentiate the activation function of neurons to compute the synaptic weights modification. Since spiking neurons are non-differentiable functions, backprop cannot be directly transposed to SNNs. Until recently, this made deep SNNs (i.e. multilayer) and recurrent SNNs hard to train and was preventing SNNs to reach equivalent performances as

deep networks. Several strategies have started to emerge and were proposing alternative to make SNNs compatible with backprop algorithms. The key idea here is to find some equivalent technic to compute the gradient descent, which involves differentiating the neuron activation function. Smoothing of the activation functions or surrogate functions to the gradient are very attracting options that have demonstrated equivalent performances of multi-layer SNNs with their equivalent ANNs. Nevertheless, these technics are approximation of gradient descent technics and are unlikely to bring SNNs above ANNs performances [6].

A very distinctive aspect of NCE with ANNs for learning is to consider more bio-inspired learning rules. A very famous example is the proposition of Spike Timing Dependent Plasticity (STDP). STDP is a variation around the seminal Hebb's idea of "who fire together, wire together" and was identified in biology. Spike timing is used to define correlation in between a pre and post-neuron (pre fires before post) and anti-correlation (post fires before pre). The correlation (anti-correlation) of activity is used to define weight potentiation (depression) during learning. A key aspect of STDP, which is not present in ANNs, is to propose a local learning rule (i.e. weight modification depends only on pre and post-neuron activity). This is highly desirable from a hardware perspective since it could enable the development of massively parallel hardware substrates where information doesn't need to travel extensively in between computing nodes. At the opposite, backprop suffers from the spatial credit assignement issue, which consists in the problem of calculating the weight modification based on some loss function calculated at the output of the network and to retro-propagate this error across it. Nevertheless, STDP was limited until recently to single layers SNNs and was not adapted to deep SNNs. A direct consequence was a poor level of performances with respect to ANNs. There has been recently some breakthroughs along this line with the proposition of neo-hebbian learning rules [7]. In these extensions of Hebb's rule, a third factor is added to the standard two factors (e.g. pre-neuron activity and post-neuron activity in STDP) and result in three-factor learning rules. For example, an additional learning signal to the STDP can be used to indicate how much the local learning is useful with respect to an objective function describing the network performances. This strategy, even if trading-off the locality of learning, was able to demonstrate high performances of multi-layer SNNs and could be a game changer in the development of deep SNNs.

From a different angle, important progresses has been realized recently toward high performances SNNs. In these approaches, the key idea is to find a way to calculate the error of the network based on local information only. In other words, the the idea is to calculate locally the backprop signal. Equilibrium propagation (eq-prop) corresponds to an energy-based model of the network and was used to derive a learning rule, which was able to backpropagate the error [8]. In eq-prop, error retro-propagation is based on two main ingredients. Weights modification is associated to a local term that depends on the local activity of the pre and post neuron, and on a global term describing the network performances. This approach is again promoting locality with respect to standard backprop algorithm. The local learning term was also demonstrated to be equivalent to some extend to STDP. Note that this proposition belongs to a broader stream that tends to bring ideas from ANNs to SNNs [9]. A second promising direction was also proposed for recurrent SNNs. Eligibility propagation (e-prop) is proposing to define learning based on a local learning term and some eligibility traces that are indicating to the network "when to learn" based on the global objective function to implement [10]. E-prop have shown equivalent performances to LSTM ANNs, which are one of the most widely used recurrent ANNs network. While important differences exist in between eq-prop and e-prop, both are offering new perspective for deep SNNs deployment since locality of learning is preserved to some extend and could be deployed on massively parallel hardware substrates.

Biology is also pointing toward a combination of local / global learning in neural network [11]. An example of this idea is the tri-partite synapse. In this synaptic model, learning depends in addition to pre and post-synaptic factors (e.g. spike activity, membrane potential,…) on the astrocyte signal that is supervising the actual learning. Astrocytes are non-spiking cells that are known to regulate calcium ions concentration (among other functions) across multiple neuronal cells. Through this calcium regulation, learning can be strengthened or weakened. Other evidence coming from biology are also strengthening the combination of local and global effects. For instance, dopamine release during learning has been demonstrated to be a

reward signal triggering synaptic plasticity. All these elements can find analogy to some extend to the previously mentioned learning rules. But it is to note that these mechanisms are involving complex spatial and temporal dynamics during learning that are making bio-realist learning algorithms hard to define.

1.4. ENERGY CONSUMPTION CHALLENGES

Energy consumption remains a major challenge for the development of AI. If ANNs algorithms have demonstrated attractive performances for various tasks, their deployment on edge applications that requires strong constraints on the hardware energy budget is still limited. ANNs are indeed data intensive systems that require massive exchange of information in between the computing nodes and the off-chip memory (these signals could be either input data or models parameters such as synaptic weights). Conventionnal hardware are mostly using DRAM memory to store data and parameters and most of the energy is dissipated through data movement in between the different element of the system. Note that this statement applies for both conventional CPU and GPU, even if the later is allowing for higher throughput (throughput being the number of operations per second, TOPS). To this end, in-memory computing (IMC) has attracted a deep interest to reduce energy consumption associated to data movement [12]. IMC is intimately linked to the hardware concept of embedded memory where the physical devices for memory are integrated along with the CMOS computing elements, thus limiting the physical distance of data movements. If IMC could be realized with various technologies (embedded DRAM, SRAM, for instance), a very attractive option is to used resistive memory technologies (RRAM or memristors). In this approach, memristors are used to implement synaptic weights and the Kirchoff's laws are used to realize the key operation of dot product (note that dot products, vector matrix multiplication and Multiply And ACcumulate are essentially the same basic operation). Benefits of this approach are two sides: (i) IMC of dot product allows for reduce latency since the operation could be realized in principle in a single time step and, (ii) data movement is limited thanks to on-chip synaptic weight. This later aspect is of first interest for implementation of on-chip learning strategies. Note that the dot product operation in itself doesn't present a significant interest in term of energy with todays technologies since memristors require to sink a significant current during writing and reading. But if IMC with memristor is still a very active research direction for ANNs implementation that could enable high throughput with low energy (TOPS/W), challenges still exisits regarding the variability of memristors, which makes them good candidates for low accuracy computing, but less attractive for conventional backprop algorithms requiring high accuracy weight updates during gradient descent.

Innovative hardware (i.e. IMC) are partially solving the energy consumption issue of ANNs but power consumption of biological systems seems still out of reach. Note that this applies for both ANNs and SNNs. Nevertheless, SNNs are pointing toward the possibility of better energy performances with respect to ANNs thanks to their distinctive data representation. Spike encoding of information is pointing toward a very efficient way of computing information if spatial and temporal sparsity is obtained. Spatial sparsity corresponds to a good distribution of the information along the different nodes of the network. Temporal sparsity corresponds to a representation of signals with as few spikes as possible. Both spatial and temporal sparsity are also favored by encoding only the essential part of the signal. For instance, when computing ANNs models involves dense vector matrix operation at each "frame" of the signal, SNNs are distributing over time the same operation, thus limiting instantaneous power consumption. A second aspect that spike computing seems to offer is the ability to compute with low accuracy and large noise, while preserving efficient performances. Indeed, biological networks are using 2-3 bits resolution for synapses and their signal to noise ratio are well behind the requirements of digital technologies. In term of energy, this have a direct impact since accuracy in electronic systems comes at the price of higher energy consumption and higher latencies.

From a hardware perspective, biology is also pointing toward key differences with current approaches. An interesting example is to compare the subthreshold voltage slope used by ionic channels in comparison to transistors channel in the seminal work of Carver Mead [13]. This important difference could be explain to some extend by multiple aspects. (i) Biology is using various carriers of information, and in particular ions, for computing. For instance, divalent ions are reducing by a factor of two the energy required to overcome the thermal barrier used for charge separation in a physical systems (this Boltzmann limit in transistor corresponds directly to the subthreshold voltage slope of about 60 mV in transitors). (ii) Electronic systems are relying on fast moving and confined electrons through drift and diffusion, which implies important Joule effects. At the opposite, biology is based on slow motion of ions in a dilute conductor (i.e. the electrolyte) mostly driven by diffusion. (iii) Separation of charges in biology doesn't rely solely on electrostatic effects but also employs mechanical components such as ionic pumps. These elements suggests that reaching the energy performances of biology with current approaches would require rethinking profoundly the choice of substrate used for computing.

1.5. THE INTEGRATION CHALLENGE

In 2010, the Synapse project proposed a roadmap for the development of brain-inspired hardware. The most critical metric was the synaptic connection since density of synapses is around 10000 times larger than neurons. They estimated the footprint of an electronic synapse to be 100 nm^2. Such ultimate footprint would ensure ultra-high integration of synaptic elements to match what is observed in the brain. Memristors have been strongly considered to fill this requirement since sub-5 nm devices has been demonstrated (note that flash technology scaling has been stopped at the 28 nm node). In addition to ultra-small footprint, memristor in its simplest version is also compatible with crossbar integration scheme, which could allow for high-density integration. Nevertheless, passive crossbar (i.e. memristor are integrated at the crosspoint of two metal wires without selector) have faced several challenges that still need to be answered. The most critical ones are crosstalk effects (i.e. undesired programming of adjacent memristors during writing of a specific one) and mismatch of the crossbar wires impedance with the memristors. This later effect worsen when scaling the crossbar dimensions and would require important effort at the technological level to decrease wire resistance and increase memristor resistance.

Another important limitation to high-density integration of memristor appears to be the complexity of the overhead circuitry required to drive the memristor elements. This overhead increase when writing scheme of the memristors becomes more complex, in particular during learning of the synaptic weights. It appears today that 1T1R integration is favored to solve the crosstalk issue. But integrating complex writing schemes to implement synaptic plasticity are requiring 2T1R for STDP implementation, and up to 6T2R1C for backpropagation [14]. A very attractive option to limit this increased footprint associated to the writing circuitry is to rely on the physics of memristor technologies to implement locally various plasticity rules and synaptic mechanisms. For instance, various drawbacks of conventional memories such as retention or stochastic switching could be turn into advantages to reproduce synaptic effects during learning and operation.

An additional difficulty associated with integration density in ANNs and SNNs is the interconnection of the different nodes in a parallel manner. Since conventional CMOS technologies are bounded to a 2D integration of the transistors, important interconnects in the middle-end of line are required. This constraint is pointing toward the necessity of 3D integration of computing elements, and in particular of the synaptic connections. There is a true potential to integrate memristors in 3D but this objective will require important technological efforts.

A more fundamental limitation in the development of computing substrates that could match the requirements of high density of neurons and synapses in ANNs and SNNs is the top-down approach imposed by conventional technologies. If general-purpose computers are offering a large abstraction between the

algorithms and the physical substrates, the development of neuromorphic circuits is creating a strong link in between both. It means that the complexity of the algorithm (number of nodes and parameters) is directly mapped on the physical substrate used for computing and need to be known before fabrication. Even if sparse networks are obtained after learning by using pruning technics for example, the initial network topology needs to be over-estimated in order to enable learning. Definition of the optimal topologies that could be used without trading too much on the flexibility is a very important question that is only partially answered today. At the opposite, biological networks are using relaxed device dimensions (i.e. neurons are about 30 um and synapses are 100-200 nm). However, they evolved following a bottom-up strategy, which implies that material resources are created when required and are used very efficiently. This aspect, in addition to the truly 3D integration of the brain could be a key ingredient that technology would need to reproduce in order to reach the level of performances of its biological counterparts.

1.6. CONCLUSION: TOWARD A TRUE ARTIFICIAL INTELLIGENCE?

NCE appears as a promising solution to bring current AI toward its next generation. The first practical aspect of bringing AI to the next level is to unlock the energy consumption challenge. This could result in the deployment of AI on a variety of embedded applications near sensors, which are regrouped into the class of edge computing applications. Having intelligent computing embedded on portable devices will reduce the bottleneck of data exchange to / from the cloud and its (often not displayed) related energy consumption. This aspect is two sides since it will reduce the energy sink of large data centers and reduce the pressure on the battery lifetime of portable devices. From the previous elements describe above, we see that the main vector for this “AI revolution” is based on hardware innovation that NCE is promoting thanks to its strong roots with hardware physics and on how moving from bio-inspired observations to hardware implementations. Nevertheless, reaching this goal needs to approach the problem globaly and to consider both data encoding (section 1.2), learning (section 1.3), energy consumption at the device (section 1.4) and system level (section 1.5).

But there is also a more profound impact of NCE on the future of AI. Today, ANNs are currently surpassing all other approaches in terms of performances for a large variety of AI tasks. This is suggesting that there is not a real need in innovation in computing, but continuous scaling of ANNs could be enough to sustain the deployment of AI. Nevertheless, ANNs are still not convincing on their ability to promote a “real” intelligence. It seems that some ingredients to engineer intelligence on hardware systems are still missing. This requires of course to define what can be considered an intelligent behavior, and to have some metrics to compare the performances of a deep network with the one of a sea cucumber or a fly. NCE is holding some promises to this end since it could promote a truly interdisciplinary research at the frontiers in between biology, neurosciences, computer science, and electrical engineering. This interdisciplinary approach could help us progressing toward the understanding of data representation in the brain (section 1.2), learning (section 1.3) and biological wetware principles (sections 1.4 and 1.5). If it is hard to identify a safe methodology toward this goal, an engineering approach would certainly benefit from material integration of the different NCE concepts toward embodiment of AI on hardware that could integrate every levels from sensing to computing. This should require sustained efforts on neuromorphic sensors development, online (even continual) learning circuits and autonomous computing systems deployment.

[1] LA BARBERA, Selina, ALIBART, Fabien. Synaptic plasticity with memristive nanodevices. In : *Advances in Neuromorphic Hardware Exploiting Emerging Nanoscale Devices*. Springer, New Delhi, 2017. p. 17-43.
[2] BRUNEL, Nicolas. Modeling point neurons: From Hodgkin-Huxley to integrate-and-fire. *Computational modeling methods for neuroscientists*, 2010, p. 161-185.

[3] BARLOW, Horace B., *et al.* Possible principles underlying the transformation of sensory messages. *Sensory communication*, 1961, vol. 1, no 01.
[4] SCHRAUWEN, Benjamin et VAN CAMPENHOUT, Jan. BSA, a fast and accurate spike train encoding scheme. In : *Proceedings of the International Joint Conference on Neural Networks, 2003.* IEEE, 2003. p. 2825-2830.
[5] LIU, Shih-Chii et DELBRUCK, Tobi. Neuromorphic sensory systems. *Current opinion in neurobiology*, vol. 20, no 3, p. 288-295 (2010)
[6] NEFTCI, Emre O., MOSTAFA, Hesham, et ZENKE, Friedemann. Surrogate gradient learning in spiking neural networks: Bringing the power of gradient-based optimization to spiking neural networks. *IEEE Signal Processing Magazine*, 2019, vol. 36, no 6, p. 51-63.
[7] GERSTNER, Wulfram, LEHMANN, Marco, LIAKONI, Vasiliki, *et al.* Eligibility traces and plasticity on behavioral time scales: experimental support of neohebbian three-factor learning rules. *Frontiers in neural circuits*, 2018, vol. 12, p. 53.
[8] SCELLIER, Benjamin et BENGIO, Yoshua. Equilibrium propagation: Bridging the gap between energy-based models and backpropagation. *Frontiers in computational neuroscience*, 2017, vol. 11, p. 24.
[9] LILLICRAP, Timothy P., SANTORO, Adam, MARRIS, Luke, et al. Backpropagation and the brain. Nature Reviews Neuroscience, 2020, vol. 21, no 6, p. 335-346.
[10] BELLEC, Guillaume, SCHERR, Franz, SUBRAMONEY, Anand, *et al.* A solution to the learning dilemma for recurrent networks of spiking neurons. *Nature communications*, 2020, vol. 11, no 1, p. 1-15.
[11] ROELFSEMA, Pieter R. et HOLTMAAT, Anthony. Control of synaptic plasticity in deep cortical networks. *Nature Reviews Neuroscience*, 2018, vol. 19, no 3, p. 166-180.
[12] Sze, V., Chen, Y.-H., Emer, J., Suleiman, A. & Zhang, Z. Hardware for machine learning: Challenges and opportunities. In 2017 IEEE Custom Integrated Circuits Conference (CICC), 1–8 (IEEE, 2017).
[13] MEAD, Carver. Neuromorphic electronic systems. *Proceedings of the IEEE*, vol. 78, no 10, p. 1629-1636 (1990)
[14] IELMINI, Daniele et AMBROGIO, Stefano. Emerging neuromorphic devices. *Nanotechnology*, 2019, vol. 31, no 9, p. 092001.

2.CHAPTER 2

In-Memory Vector-Matrix Multiplication in Monolithic CMOS-Memristor Integrated Circuits: Design Choices, Challenges, and Perspectives

2.1. INTRODUCTION

The semiconductor technology sector, and particularly its research core, are currently undergoing fundamental changes. After decades of predictable evolution based on the strategy relying on CMOS scaling[1] yielding incremental processor performance improvements, new solutions are required[2]. The first driving force for this revolution is energy consumption, which remains a major challenge for the ubiquitous deployment of electronic chips on an ever-increasing number of devices[3]. Solving this challenge would enable both: the integration of more computing functions on a variety of portable miniaturized devices with demanding energy/form-factor constraints, and more generally, conserving the total energy required to power billions of electronic devices. The second driving force is the massive deployment of artificial intelligence (AI) in our everyday life, which is redefining the basic principles of the hardware architecture required for computing. In particular, von Neumann computing architecture[4] is not well adapted to machine learning (ML) implementation, which is a main vector for the widespread adoption of AI. Indeed, implementations of ML algorithms on standard CPUs are typically inefficient in term of speed due to the constant dataflow between arithmetic units (AUs) and the memory, limited by the von Neumann bottleneck. There is consequently an important need to improve computing efficiency from both an energy consumption and throughput perspective. To this end, hardware innovation is expected to play a major role by offering viable solutions to sustain the deployment of electronics.

Specialized hardware such as GPUs[5], which are highly parallelized versions of classical von Neumann CPUs, have been game changers in the acceleration of ML. However, they are offering only a partial solution to the speed and energy challenges. More precisely, GPUs are a first step toward hardware specialization where the key operation of Multiply and Accumulate (MAC) has been parallelized in order to offer important speed improvements. Since MAC operation represents the most intensive calculation required for ML algorithm implementation, it explains why GPUs have led to important breakthroughs in acceleration of ML by enabling training and operation of deep neural networks[6] in a reasonable amount of time. But parallelization solely cannot solve the energy challenge for two reasons: (i) intensive data movement between the different physical elements of the hardware results in important energy consumption (i.e. data movement between on-chip memory and AU, but also data movement in between the different on-chip and off-chip memory level) [7]; (ii) as in

CPU, the fundamental algorithmic operation is still realized with the same elementary logical operations, which require the same energy budget.

Improving both energy and speed requires to rethink more deeply hardware design principles and prudently explore emerging computing technologies. Along this line, more advanced solutions exploit hardware specialization even further and propose to design application processing units (APUs), which optimize the throughput and energy requirement for a specific application (Figure 1(a)). In these approaches, innovation is more supported by hardware diversification and specialization, rather than by software innovation to make a balance between their functional flexibility and performance [2]. By deploying hardware specialization, there have been several low power research chips, data center chips and cards proposed in addition to recent advancements in CPUs and GPU-based neural engines. However, it should be noted that reaching an end-to-end solution for an efficient hardware will require scrutinizing other computing paradigms and technologies. In this context, in-memory computing architectures enable efficient computing with negligible data movement by co-locating memory and processing unit. This path has been explored with various technological solutions, from mainstream SRAM and DRAM to more emerging ones such as eDRAM [8]. Beyond charge-based digital memory technologies, in-memory computing based on non-volatile resistive switching devices (RS) monolithically integrated on CMOS is opening new perspectives for ultra-efficient MAC operation engine development [9]. Firstly, monolithic integration of memory in close vicinity of logical units reduces significantly the distance for data trafficking, and thus should reduce energy consumption and throughput limitation[9-11]. Secondly, in-memory computing represents a new physical implementation of the basic MAC operation with potential for important improvements with respect to the same criterions.

In this paper, we aim to review the main limitations and opportunities of in-memory computing with resistive memories for MAC operation engine, also known as Vector Matrix Multiplication engine (VMM engine). On this basis, as shown in Figure 1(b), the challenges hindering toward the path of monolithically integrated resistive memory and CMOS VMM engines to become a mainstream hardware have been categorized into three different levels: physical constraints, circuit-level challenges and system-level challenges. Initially, we define the main issues corresponding to physical limitations of this specific class of hardware e.g. accuracy, integration, scalability and speed. Subsequently, we assess the circuit-level challenges and analysed the input and output circuit design costs and opportunities. Finally, system-level obstacles such as data movement and data conversion issue have been discussed. Also, we propose a rational analysis of such APUs performance and their trade-offs in the context of ML applications, but the same reasoning could be applied to a wider range of applications [12] such as image processing[13,14], combinatorial optimization [15-19], sparse coding [20,21], associative memory [22-26], deep learning inference/training [27-30], unclonable functions [31-34], principle component analysis [35,36], spiking neural networks[37-41], solving linear [42], and partial differential equations [43] and reservoir computing [44-46]. Our intent is to provide a comprehensive analysis to assess the novelty of the reviewed examples and discuss different design choices to better understand this emerging class of hardware and to rationalize performances evaluation.

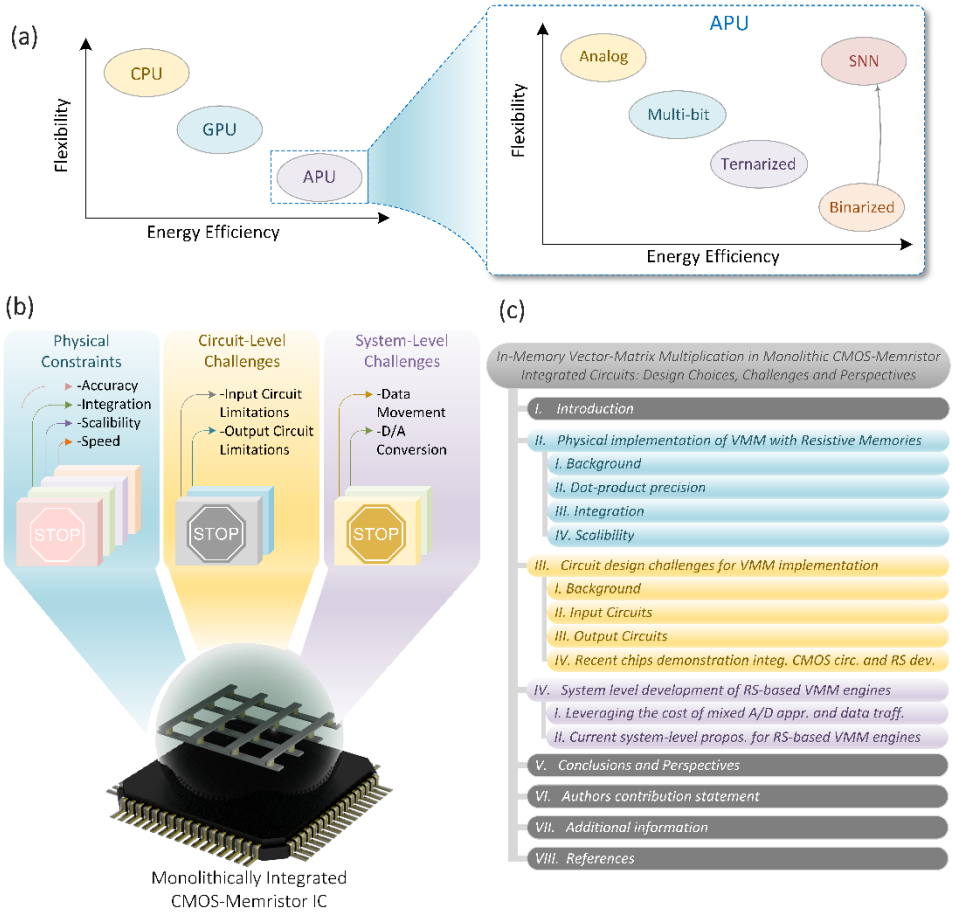

Figure 1: Different computing hardware performance overview and challenges and limitations hindering the path toward of monolithic CMOS-memristor VMM integrated circuits to become mainstream AI hardware. (a) Here, a simple view of APU platform's energy efficiency performance and its flexibility in terms of application range is compared with conventional platforms like CPUs and GPUs. Different RS-based APU classes with low to high resolution weight networks are displayed in terms of energy efficiency and application spectrum flexibility. At the opposite to the trade-off between flexibility and energy that current hardware are experiencing, Spiking Neural Networks (SNN) observed in biology combine both flexibility and low energy consumption. Finding the keys for this implementation seems a disruptive direction for future hardware design. (b) The challenge has been divided into three different categories: physical constraints, circuit-level challenges and system-level challenges. (c) Manuscript's tree structure.

2.2. Physical implementation of In-Memory computing with RRAM

2.2.1. Background

VMM is the main operation module required to implement neural network structure (Figure 2(a)). The first basic function required for VMM physical implementation is the multiplication between two real numbers a and b ($a \times b = c$). In digital logic, multiplication is realized by pipelining multiple full-adders (Figure 2(b)). The precision of the multiplication is defined by the digital representation of the real numbers (number of bits, floating/fixed point). Resistive memory on the other hand offers a new concept for implementing multiplication leveraging Ohm's law where the current I is equal to voltage V multiplied by conductance G ($V \times G = I$) (Figure 2(b-c)). The advantages of this approach are two-fold: (i) only a single time step is required to compute the multiplication versus multiple time steps in digital implementation and (ii) energy consumption is considerably lower.

Projected resistive memories performance for an average resistance of R = 1 MOhm, read voltage of 0.1 V with pulse duration of 1 ns, the energy consumption equals $E_1 = 0.1\times10^{-7}\times10^{-9} = 10^{-17}$ J. Note that with todays performances, the energy calculation should consider R = 10-100 kOhms, V = 0.1 V and t = 1 µs leading to $E_2 = 0.1\times10^{-5/-6}\times10^{-6} = 10^{-12/-13}$ J. These energy consumption should be compared with 8-bit digital multiplication of E_3 = 0.2 pJ with 45 nm CMOS technology node~\cite{horowitz20141} pointing out the important gain attainable only if resistive memory improvement is sustained.

The second basic operation required by VMM is the addition. While this operation is carried out by adders in digital electronics, this can also be implemented physically in the analog domain by summing all currents resulting from each multiplicative element in a shared metal line (Kirchoff's law). This strategy shows a clear advantage for speed improvement due to its highly parallel manner as the Add operations are carried out within multiple parallel channels of the crossbar simultaneously in a single clock cycle with the multiplications. For the sake of comparison, one 8-bit full-adder uses approximately 200 gates in conventional CMOS design and requires number of computing cycles proportional to the Add operation precision. These two basic multiplication and addition operations correspond to the fundamental MAC operation or dot-product, which constitutes the core of VMM.

While this qualitative analysis highlights the advantages in terms of speed and energy consumption of in-memory computing for VMM engine implementation, a fair comparison with digital CMOS technology is more complex and limitations start to appear due to non-ideal parameters such as physical constraints, overhead circuit design and system level operation.

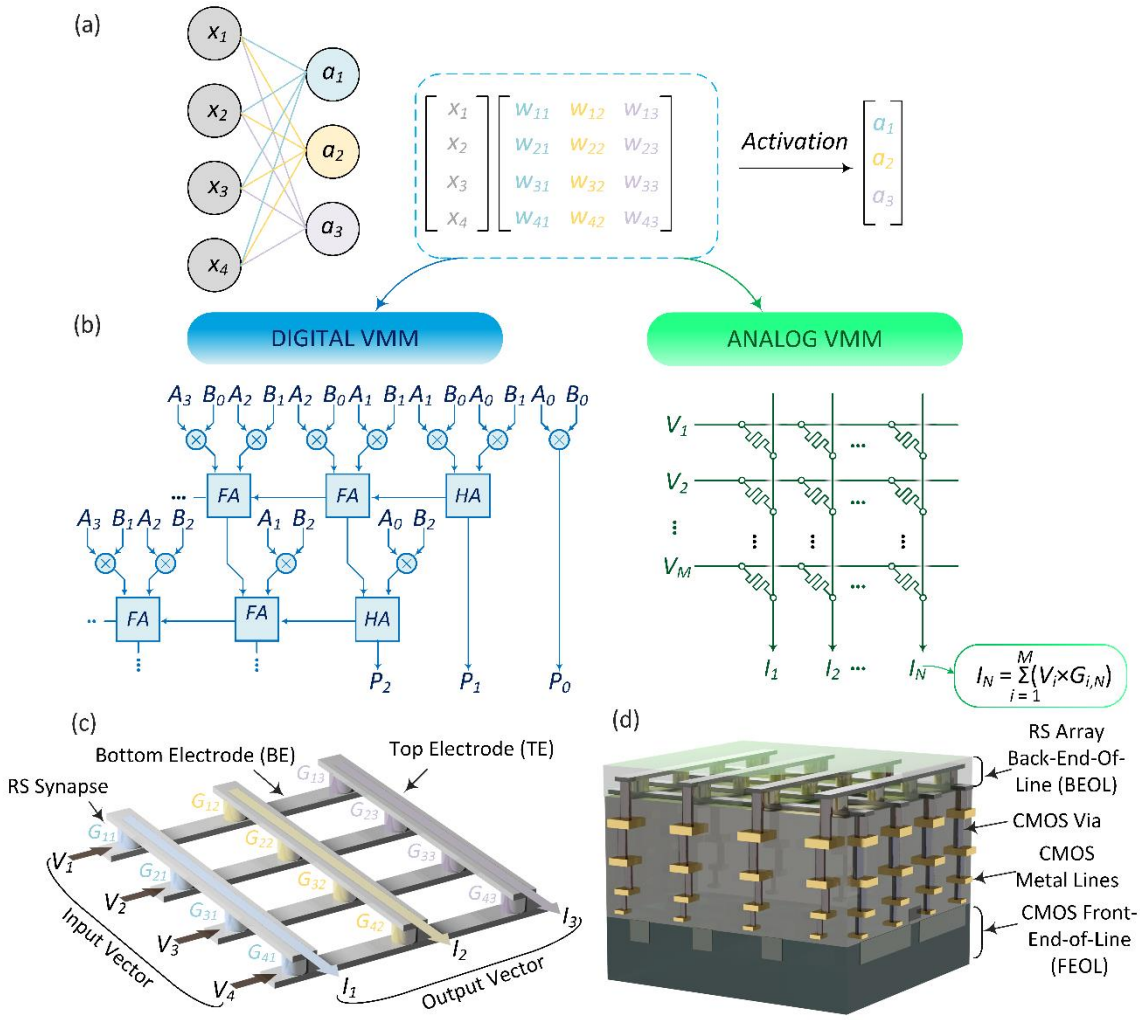

Figure 2: (a) A basic neural network structure is shown including input vector, weight matrix and output vector. (b) Schematic of the digital and analog vector matrix multiplication and their implementations. VMM digital implementation is realized by pipelining multiple adder and multiplier digital blocks. Analog VMM on *M*×*N* RS-based crossbar is realized by summing currents from *M* lines in *N* columns. (c) The physical implementation of RS-based VMM engine shows input vector is applied as a voltage vector into the world-line of the array (bottom electrode), weight matrix is stored on RS device conductance and output is sensed as accumulated current in the bit-line (top electrode). (d) 3D illustration of RS-based crossbar monolithically integrated on top of the CMOS using back-end-of-line (BEOL).

2.2.2. Dot-product Precision

Resistive switching (RS) devices have been developed following two main research directions. On the one hand, resistive switching mechanism has been investigated as a potential solution for the development of a universal memory. This kind of binary memory, called Resistive Random Access Memory (RRAM), could combine high switching speed (sub-ns), low energy (pJ range) and high endurance (10^{12} cycles) of DRAM and SRAM with non-volatility (>10 years retention) and scalability (<10 nm) (Figure 3(a-b)). Various RRAM cell candidates, among which HfOx and TaOx RRAM are the best representatives (figure 3c), are already integrated in fabrication lines of industry and integrated with CMOS technology[47]. They take advantage from CMOS technological maturity and reliability and have been exploited mostly in digital applications such as storage class memories (i.e. Flash). Some recent works have investigated the possibility to store few discrete conductance levels in a single memory cell resulting in up to 3-bit multi-level cells. This kind of device can either implement a 1-bit dot-product or a low resolution, e.g. < 3-bit dot-product [48].

On the other hand, many research groups have focused on resistive switching mechanism for memristor or memristive device implementation (Figure 3d). The association between the theoretical concept proposed by Chua [49] and a possible physical implementation of this new circuit element [50] has open new perspectives for circuit design, and specially for VMM. In the ideal memristor framework, resistive switching is used to implement a variable resistor where continuous resistive states can be reached by controlling the voltage (or current) applied to (through) the switching material. In that scope, the number of conductance states that can be stored in the memristive element directly defines the precision of the in-memory dot-product computation. In recent years, optimization of memristive device has focused on the resolution and controllability of the analog switching using various switching mechanisms and materials such as transition metal oxides, ferroelectric tunnel junctions or more exotic materials (See [51] for a review of the different options). Memristive devices have demonstrated analog switching controlled by analog pulses of voltage equivalent to 8-bit accuracy, paving the way for 8-bit dot-product [52]. The 8-bit accuracy has been demonstrated on discrete devices and only 4- to 5-bit resolution has been reported for integrated devices due to parasitic effects induced from other circuit elements[53].

The maturity of memristive technologies is not as developed as the RRAM technology, which results in inferior performance regarding endurance, retention and speed. There are still several research opportunities in this area and efforts need to be pursued to improve memristive devices overall performance. However, there is currently no strategy nor materials allowing to reach the 32-bit dot-product precision offered by digital approaches. This imposes limitations in terms of VMM applications, such as deep neural networks that relies deeply on high accuracy calculation of the synaptic weights during training [54]. In that scope, innovations in integration schemes could greatly improve the accuracy of the memristor-based VMM. For instance, while RRAMs differ from analog memristive devices by the difficulty to access to intermediate resistance states, there is, in principle, no physical limitation to have multi-level analog states in RRAM. HfO_x-based RRAM, usually exhibiting sharp SET and semi-gradual RESET[55], can be better controlled by using analog current limitation mechanism through an access transistor to implement analog switching close to 5-bit precision[56]. The trade-off here is between a more complex cell design and a higher precision of programming. Along this line, one interesting approach proposed by [57] consists in a hybrid architecture, where two Phase Change Memories (PCM)resistive cells are coupled with six transistors and one capacitor (1C6T2R). Small weight increments, or decrements, are accumulated on a capacitor and stored back in the non-volatile resistive element once accumulated changes fall within the resolution range. Such integration widens the range of VMM applications like in-situ training while decreasing energy consumption compared to contemporary von Neumann architectures. This resolution improvement comes at the cost of more complex resistive cells design and additional shared control circuitry. Short- and mid-term efforts should be dedicated to more complex resistive cells design that would leverage design complexity with controllability and precision for analog VMM implementation.

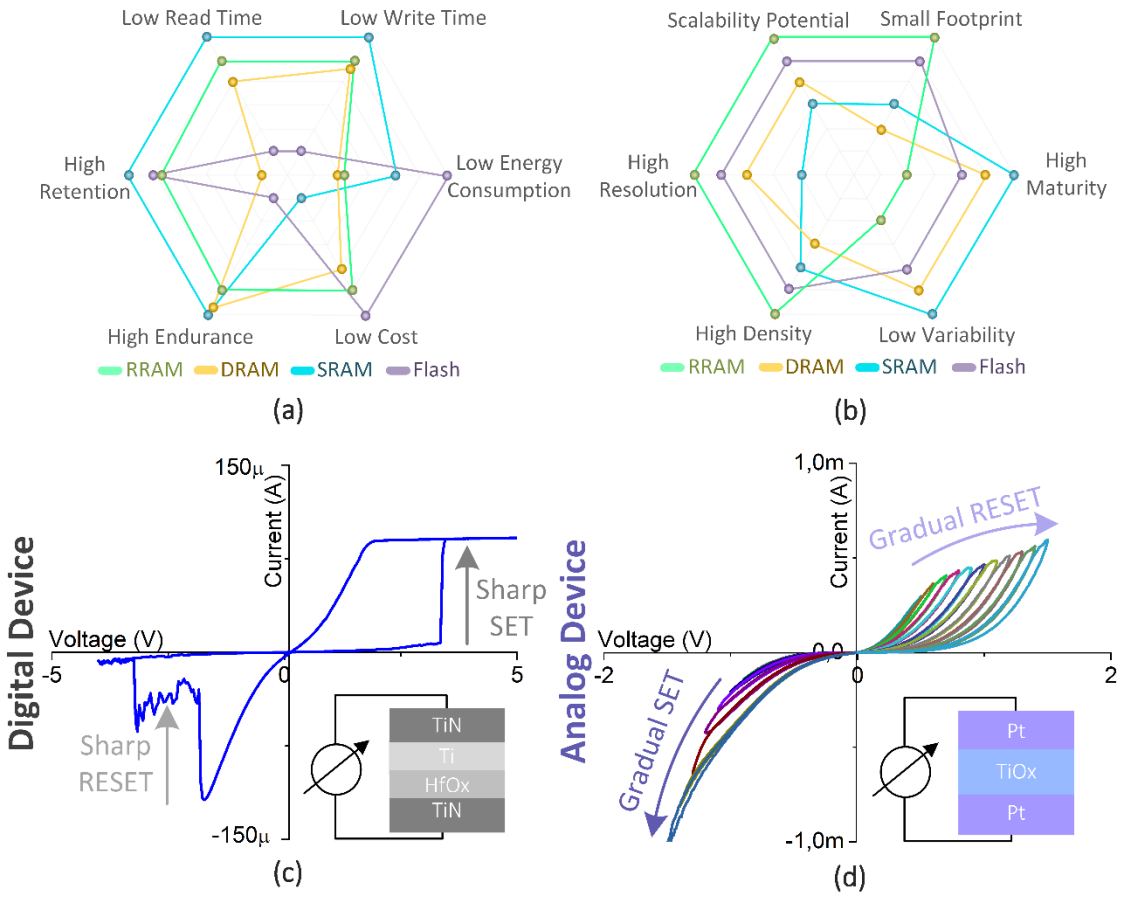

Fig. 3. (a) In this spider diagram, RRAM, DRAM, SRAM, and Flash memories are compared in terms of the cost, read time, write time, energy consumption, endurance and retention. (b) In this diagram, RRAM, DRAM, SRAM and Flash memories are compared in terms of other criteria: flexibility, footprint size, maturity, density, variability and potential of the scalability. (c) The *i-v* curve of the prototypical RRAM digital HfOx and its sharp switching behavior in SET and RESET regions are depicted. (d) The switching behaviour for the prototypical memristive TiOx analog device is displayed [58]. Both RRAM and memristive devices belongs to resistive memories family and are used to construct the spider diagrams.

2.2.3. Integration

One of the substantial advantages of RS devices is their advanced integration potential thanks to their excellent scalability. Sub-10 nm switching crosspoints have been reported in[59] and[60], paving the way to surpass the scaling limitations of Flash and DRAM. In addition, the two-terminal structure of RS devices enables ultra-dense integration in crossbar arrays, in which a memory device is located at each intersection between two metallic wires resulting in a matrix-like organization. Finally, RS devices and crossbar arrays can be fabricated with CMOS high-volume manufacturing processes and materials allowing monolithic 3D integration in CMOS BEOL. This ideal approach (see Figure 2(d)) results in a $4F^2$ footprint for a single memory crosspoint, F being the critical dimension of the metal line interconnect). Monolithic 3D BEOL integration of resistive memories presents a major advantage compared to other on-chip memory technologies such as SRAM, which requires a footprint of ~$200F^2$ in the front-end-of-line (FEOL). This very attractive approach could relax CMOS scaling requirements by providing additional integration opportunities in the vertical dimension. In addition to BEOL attractiveness, the possibility to stack multiple crossbars on top of each other has been demonstrated experimentally and could be conveniently integrated with CMOS for ultra-high-

density memory circuit design[33,61]. There are still important engineering challenges to address in order to bring these concepts to their full potential: (i) compatibility of advanced lithography steps with BEOL metal layout, (ii) impact of monolithic 3D fabrication processes on the performance of previously fabricated devices, (iii) process homogeneity and yield ensuring high-quality fabrication for each layer and (iv) high-conductivity interconnects even for ultra-fine pitch. While crossbar architecture offers a truly parallel organization that could map directly the VMM operation, the main limitation comes from the difficulty to access individual memory cell accurately. Parasitic sneak path, currents coming from other resistive cells in the array, are preventing an accurate reading of each resistive element individually.

RRAM and memristive devices can be addressed with or without the use of a selector. On the one hand, RRAM requirements have favored optimizations towards accessibility and controllability of individual memory cell by adding a selector, usually a FEOL transistor, in series with the two-terminal element leading to 1T1R cells. This solution requires a transistor per memory cell with the allocation of additional silicon area and interconnects for memory management, decreasing the attractiveness of two terminal resistive memory. The resulting integration scheme is then only considered as a pseudo-crossbar array. Two-terminal selectors, such as threshold switching elements or non-linear diodes, are today attracting lots of attention toward 1S1R cells. Those passive elements can prevent sneak path currents and preserve two-terminal interconnection of each memory cell [62]. Still, 1S1R integration is facing important challenges such as (i) large variability coming from the selector itself and (ii) shorter endurance in the case of switching selectors that needs to be switched for each read operation. Detailed review in this topic can be found in[63].

On the other hand, memristor-based approaches for physical VMM have favored the concept of selector-less passive crossbar integration. While RAM operations require precise access to individual memory cell, memristor-based dot-product is different since this operation is not affected by sneak paths (e.g. all lines and columns are polarized at the same time and all resistive cells are read at the same time). More exploratory in-memory computing paradigms such as neuromorphic computing, or bio-inspired spiking neural networks, can also take advantage of a similar principle. The trade-off being to favor parallelism and aggressive integration at the cost of less accurate access to individual crosspoints sequentially. It should be noted that practical integration of crossbar on chip still requires access transistors at the *N* input lines and *M* output columns of the crossbar thus leading to (*N*+*M*)T(*N*×*M*)R actual footprint on silicon. There is consequently a strong interest in improving passive crossbar dimensions above the 64 × 64 size report so far[53].

2.2.4. Scalability

In digital approaches, computational scalability of the Add operation is ensured by pipelining simple logical operations of single bits, thus allowing for very large vector-matrix manipulation (adding multiple dot-product, for instance). The digital approach is based on a trade-off between scalability of the operation, and computing time (e.g. how many clock cycles and basic operation are required). In RS-based Add operation, adding multiple dot-products is realized in a single time step. This advantage comes at the price of higher instantaneous power requirements. Adding currents from multiple dot-products results in a large current summation that could become a bottleneck for the VMM operation (Figure 4(a,c)). Adding infinite size of dot-products results in infinite time in digital scheme while it results in infinite power for Kirchoff's law-based approach. Practically, memristor-based VMM has been reported for matrix size of up to 128×64 [14]. While this was demonstrated with pseudo-crossbar having micron size electrodes, such limitations in matrix size should become a

serious computational scalability challenge with electrodes in the tenth of nanometer range that would prevent sinking large currents through them. 64×64 VMM operation was demonstrated in[53] using purely passive crossbar with a more advanced patterning process (<200 nm). Dot-product demonstration with other integrated approaches[76,77] are today limited to small vectors dimensions, with vector dimension below 25, and impose restrictions on the VMM application. There is a concern that this limitation will get worse by decreasing the metal linewidth and will require high aspect-ratio lines to achieve a high conductivity interconnect[60]. Alternatively, increasing the mean resistance of RS devices would increase scalability significantly by reducing power consumption at the cost of lower VMM operation speed. The inference operation speed is determined based on the delay induced from the input circuits, RS-based crossbar array and output circuits. In very large RS arrays, there are several parameters which should be considered to determine the delay such as interconnect resistance, interconnect capacitance, RS cell resistance, overhead circuit's impedance and capacitance. The inference delay is calculated based on the Elmore delay model as follow,

$$t_{inf} = t_{settling} + 2.2 \times \sum_{\tau=1}^{4} \tau_i$$

Where $t_{settling}$ is the settling time of the output circuit. As it can be seen in Figure 4(b), the parameters τ_1, τ_2, τ_3 *and* τ_4 are the delays from row, RS cell, column and output circuit, respectively [78]. By considering the LRS resistance of the device much larger than the interconnect resistance between each two adjacent cells, the delays τ_3 *and* τ_4 are dominant in very large arrays. By increasing the LRS of the RS cell, the inference time delay increases as it is impacting both τ_3 *and* τ_4. Therefore, the throughput of the system will be reduced accordingly. However, increasing the size of the array would also impact the inference delay e.g. increasing the number of rows will make τ_3 the dominant term to impact the total delay and it will reduce the delay. On the other hand, increasing the number of columns will increase the latency. Crossbar and pseudo-crossbar scalability challenges can also be related to computing performance (e.g. accuracy). Unlike digital approaches where input digital signals margins allow to cope with noise and parasitic, analog VMM implementation accuracy is negatively impacted in the case of large vector operations. The resulting mismatch between the resistance of the memory cells and the one of metal interconnects becomes critical in large crossbar arrays (figure 4(d)). The same bias applied on the word-line is seen differently by each cell in the crossbar due to linear voltage drops which leads to a decrease of accuracy for the VMM operation. A straightforward physical solution to these constraints is to limit the size of the crossbar array and thus the VMM performed in one step. Note that small VMM dimensions are largely used for convolutions in Convolutional Neural Networks (CNN). In conclusion, scalability of memristor-based VMM operation represents a future research direction that requires innovative solutions at both technological and system levels.

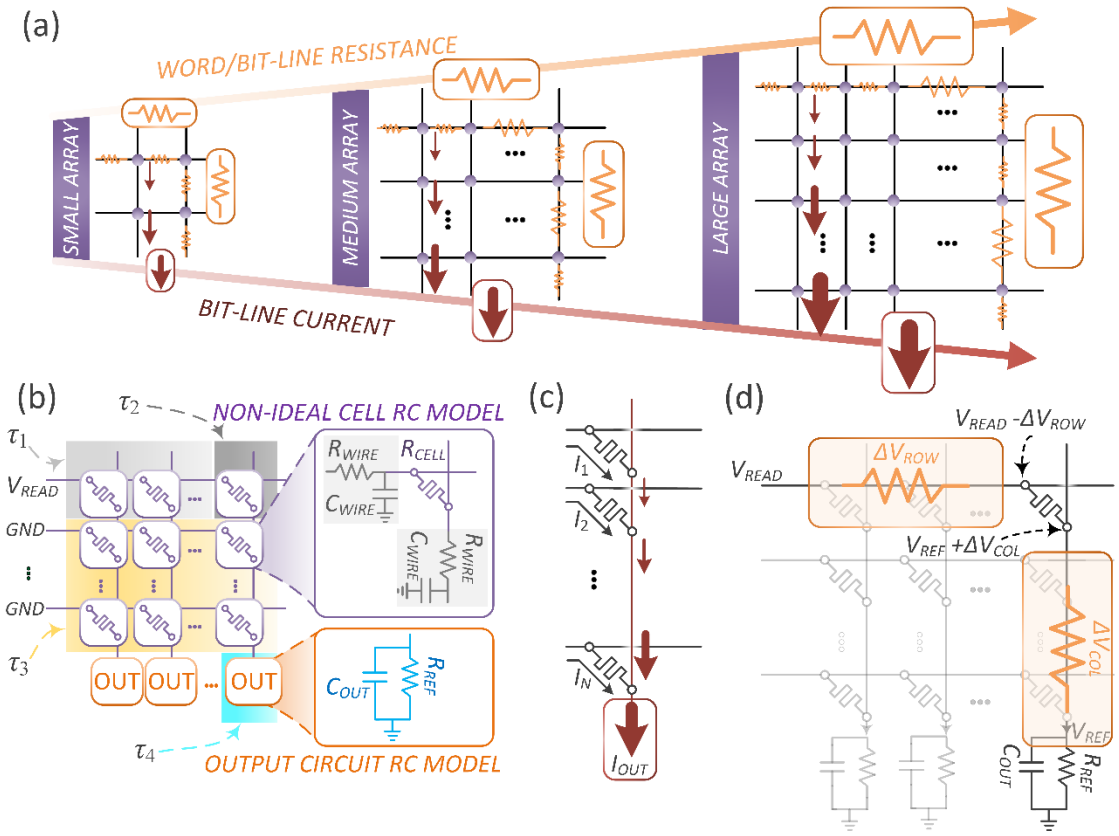

Fig. 4. Scalability challenges and RC tree Elmore delays model for RS crossbar array is displayed. (a) The scalability challenges including large bit-line current and word/bit-line resistance has been shown to get worse by increasing the size of the RS crossbar array. (b) RS crossbar RC Elmore delay model is displayed by dividing the array delay into four regions corresponding to τ_1, τ_2, τ_3 *and* τ_4 are the delays from row, RS cell, column and output circuit, respectively. (c) Increasing the number of rows increase the accumulated current in the column and could become a major bottleneck for output circuits design. Same limitation applies for large number of columns requiring to inject large current into the row and affecting input circuits design (d) The line resistance is another challenge for scalability of RS-based array due to the voltage degradation in the rows (ΔV_{ROW}) and columns (ΔV_{COL}) that can be leveraged by engineering optimization and/or compensated from input/output circuits strategies.

---------- BOX1 ----------

Addressing non-ideal parameters of RS-based system which impacts neural network accuracy

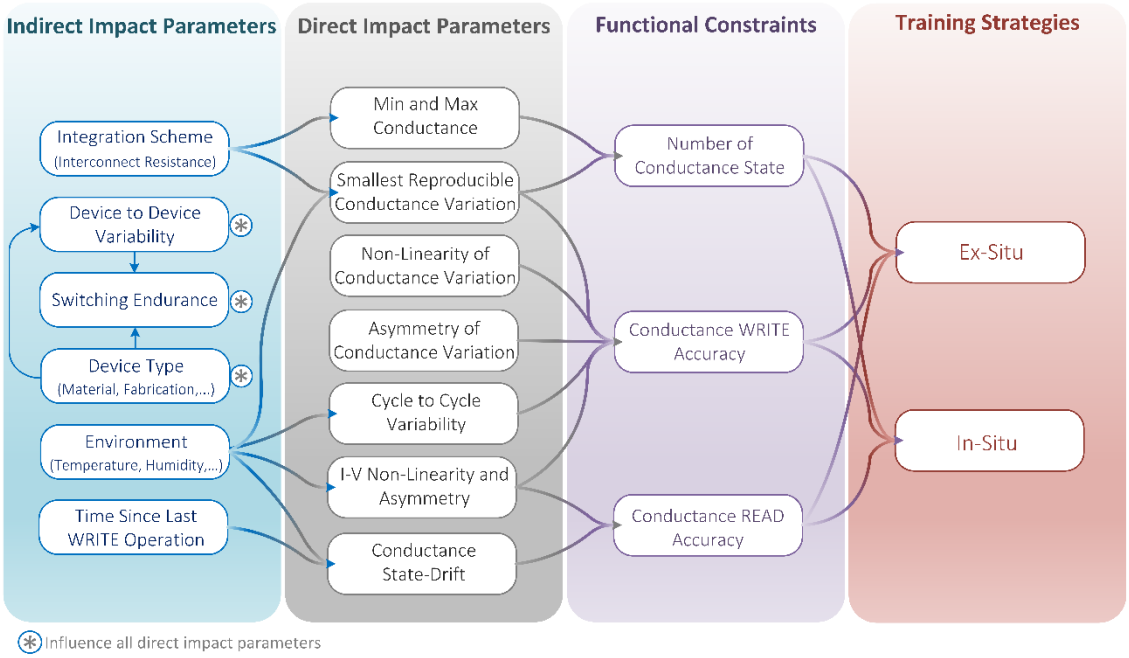

Fig. 5. Schematic classification of memristor-based system's non-idealities according to the way they influence Artificial Neural Network (ANN) accuracy. Each arrow connection should be read as "could have a significant influence on" but with no consideration for their relative impact level. The first column "Indirect Impact" can be considered as hyper-parameters that only impact the ANN accuracy through their influence on other parameters. The second column "Direct Impact" represents the fundamental parameters that directly influence the ANN accuracy. The third column "Functional constraints" lists some measurements that are often used as reference to quantify a memristive device performance.

Designing a RS-based system compatible with established microelectronic industrial technologies and large-scale production is only one part of the challenge. Since RS devices have inherent physical imperfections ~\cite{wang2019cross,adam2018challenges,Sung2018}, it is necessary to find efficient ways to deal with them. The impact level of such non-ideal parameters can be varied on different applications and here we focus on how they influence VMM-based ML applications, specifically, the accuracy of physically implemented Artificial Neural Network (ANN).

The accuracy of an ANN denotes the output success rate for a task for which it has been trained. For example, the accuracy of digit recognition using the MNIST database corresponds to the proportion of correctly classified image from a test dataset. In the context of RS-based ANN, we can distinguish two training strategies: *in-situ* and *ex-situ* ~\cite{alibart2013pattern}. In the *in-situ* scheme, the training is performed directly on the hardware by updating weights (i.e. the conductance of all devices) after each training epoch. This approach is notably impacted by all device non-ideal parameters that affect the conductance writing accuracy ~\cite{pan2020strategies,chen2015mitigating,hu2018memristor} (Figure 5) because this operation is repeated several times during *in-situ* training. In the case of *ex-situ*, the weight matrix is initially calculated in software ANN before to be transferred to the device array by encoding the determined

weights into the conductance for each cell. In that scope, the conductance programming process occurs only one time per device, which make it viable to apply advanced methods to mitigate non-ideal parameters related to writing ~\cite{pan2020strategies,alibart2012high}. Finally, a hybrid strategy showed some interesting results by fine tuning the network weights after the transfer ~\cite{yao2020fully}.

To better understand the different impacts of RS-based system non-ideal parameters on training strategies, it is interesting to not only consider their impact on functional constraints (write/read accuracy, latency, energy consumption…) but also the inter-dependence between the different parameters. For example, the *switching endurance*, which represents the average number of cycles before losing resistive switching behavior, directly impacts *minimum and maximum conductance* values over cycles ~\cite{lee2010evidence}, which in turn contribute to determine the total *number of conductance state*. Therefore, poor *switching endurance* could indirectly lead to low number of conductance state, or even failure such as stuck-at-fault where only one conductance state exists ~\cite{xia2018fault}. The impossibility to update the conductance decreases the ANN accuracy ~\cite{li2018efficient}, even more so for *ex-situ* training where weights are supposed to be mapped on working devices. The same analyse can be made with the *device to device variability* parameter, which become a problem only if this variability concerns critical device characteristics like *cycle to cycle variability* ~\cite{adam2018challenges} or the overall *asymmetry of the conductance variation* ~\cite{pan2020strategies}.

Further work should be conducted on the interactions between all non-ideal parameters in order to clarify their direct and indirect impact on the accuracy of physically implemented ANN, which could help the design and demonstration of mitigation strategies.

2.3. CIRCUIT DESIGN CHALLENGES FOR VMM IMPLEMENTATION

2.3.1. Background

As mentioned previously, projected energy consumption for a single dot-product operation can indeed be as small as 0.01 fJ, while 0.2 pJ are consumed with 8-bit digital VMM based on 45 nm CMOS technology node~\cite{horowitz20141}. However, this comparison is not a complete picture since it does not consider energy consumption for input/output signals generation. A more rigorous evaluation of memristor-based dot-product energy consumption should be done by considering 8-bit digital-to-analog converter (DAC) at the input and 8-bit analog-to-digital converter (ADC) at the output where both components consume approximately 0.1 mW and can be run at the frequency of 1 GHz (1 ns clock cycle). The total energy required to compute the 8-bit dot-product with RS devices becomes largely dominated by these DAC/ADC-based overhead circuits since $E_{DAC} + E_{ADC} = 2\times0.1\times10^{-3}\times10^{-9} = 0.2$ pJ. This simple example therefore highlights the importance of the overhead circuitry in the assessment of VMM engine performance. While most of the approaches so far have been using software-emulated or custom printed circuit boards (PCB), there are recently only a few fully integrated chip demonstrations. These demonstrations benefits are two sides: (i) exploring CMOS design overhead circuits and their compatibility with RS devices and (ii) exploring various strategies at the system level for building a fully operational chip. These choices are defining the application field of the VMM engine and impacting both the energy and accuracy performances.

2.3.2. Input circuits

VMM engines are mostly envisioned to boost energy and speed performances of conventional hardware (CPU and GPU) for specific tasks such as image compression, machine learning algorithms, combinatorial optimizations or solving linear and partial differential equations. In these applications, the VMM operation needs to be integrated into a digital environment used to manage the higher order functions such as data management and VMM definition/programing. Generating an analog input voltage from digital input data can be implemented with Digital-Analog Converters (DAC) which implies a trade-off between DAC's resolution and energy consumption. Since dot-product operation is limited to 8-bit by the RS conductance available states, there is no interest in using DACs with resolution higher than 8-bits. However, using high resolution DAC circuits will result in higher cost and reducing area and power efficiency of the VMM platform. For RS-based VMM engines, the foremost parameters used for describing the performance of the DAC are area, power consumption and more importantly the output impedance as it limits the number of memristors that one DAC can drive. In other words, the maximum output current is bounded by the DAC output impedance for a given voltage supply. The following paragraph describes an analysis method regarding the trade-off among essential DAC parameters for VMM engine applications. This method analyzes the design trade-off of a high-resolution digital-to-analog converter (DAC) with low output impedance, which is a resistive DAC with an operational amplifier (OP-AMP) follower output stage. A similar approach can be used for estimating the design trade-off among bandwidth, resolution, die-area, and power consumption for a DAC with a different architecture.

The most power-hungry blocks in the DAC are: (i) the analog circuitry that is used for driving the memristor devices, (ii) the digital circuitry that is used for storing the data and distributing the clocks. The power dissipation of the DAC can be roughly divided into the switching/leakage power of the digital circuit, and the static/dynamic power of the analog circuits. The power dissipation of digital circuits can be estimated by,

$$P_D = f_{2b} C_p V^2 + P_{Leakage}$$

where f_{2b} is the DAC maximum output frequency that equals twice the bandwidth, C is the total parasitic capacitance, V is the supply voltage and $P_{Leakage}$ is the leakage power that depends on technology node (around several pico-Watts for an inverter in 65nm technology from 1V power supply). For resistive DAC the main analog power is from the OP-AMP follower output stage, which usually employs a class-A output stage that has a maximum power efficiency of 50%. Also, the analog power can be estimated using,

$$P_A = n \times V^2/R,$$

where n and R parameters are the number of devices which are biased by the DAC and RS device resistance, respectively. Assuming the minimum resistance of each RS device is 50 kΩ, and the power supply voltage is 3.3 V, the estimated total power consumption is shown in the Figure 6(a). It has been shown that the power consumption is almost proportionate with number of devices below 100 MSample/s operating frequency and this is because the analog power is dominating when the quantity of RS devices becomes relatively large. While in higher operating frequencies

than 100 MSample/s the power consumption will be impacted mainly by operating frequency rather than number of devices when digital power becomes a dominant term.

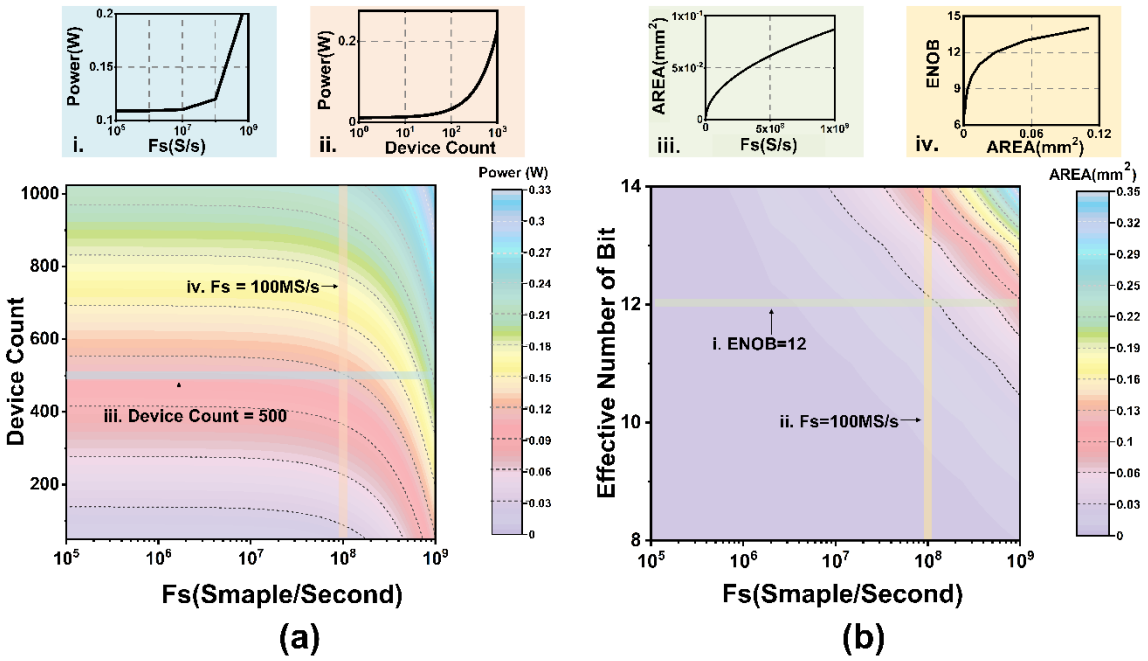

Fig 6. (a) The DAC area usages versus the sample frequency and effective number of bits. In the following two specific cases has been described by sub figures. (i) Area usage versus sample frequency at effective number of bit (ENOB) equals 12. (ii) Area usage versus ENOB at sample frequency of 100 MS/s. (b) The DAC power consumption versus sample frequency and device count (the number of memristor devices driven by one DAC), assuming the total parasitic capacitance is 10 pF. In the following two specific cases has been described by sub figures. (iii) The power dissipation versus sample frequency when the device count is 500. (iv) The device count versus power consumption at sample frequency of 100 MS/s.

The die-area is mainly constrained by the need DAC resolution. The die-area is mainly constrained by the needed DAC resolution that limited by the element matching and noise. For resistive DAC, the major noise is from the amplifier at the output stage, and input-referred noise is given,

$$V_{rms}^2 \cong \frac{4K}{C_{ox}WL} \ln\left(\frac{f_2}{f_1}\right)$$

where W and L are the width and length of the input pairs, respectively. Parameter K is Boltzmann's constant, C_{ox} is the gate capacitance per unit area, and f_1 and f_2 are the low corner and high corner frequencies, respectively~\cite{carusone2011analog}. The matching of the resistor is described as follows,

$$S_R = \frac{1}{W_R\sqrt{R}}\left(k_a + \frac{k_p}{W_R}\right)$$

where W_R is the width of each resistor and R is the resistance, and k_a and k_p are the constants that highly depend on the technology representing the contributions of area and peripheral fluctuations [80]. Figure 6(b) shows the estimated area of resistive DAC versus the operating frequency and the

effective number of bits. The area is changing almost linearly with the operating frequency and exponentially with the effective number of bits. Similar approach can be used for estimating the area and power consumption for the DAC with different architecture.

In addition to undesirable high energy consumption of the high-resolution DACs, delivering perfect analog input signal on each memory cell is challenging since it can be easily deteriorated by crossbar arrays imperfections. As mentioned previously, voltage drop along the metal lines (Figure 4(c)) induces analog values distortion (each resistive memory from a line will be subjected to analog voltage drops when the distance from the input circuit increases). This issue can be solved by additional computing overhead via software processing of the data as proposed by[72]. In this approach, voltage drop along the metal lines is calculated and compensated by RS conductance adjustment. Another limitation affecting the VMM accuracy when the input data is encoded with the analog voltage amplitude signals is the non-linearity of the current-voltage characteristic of RS elements. In this case, the actual conductance of the RS element is input-dependent and can impact the VMM resolution. This problem could again be tackled by data pre-processing including the effect of RS devices' non-ideal parameters into the analog input but can become quickly very complicated if high variability in RS device is to be integrated in pre-processing. Alternatively, [76] proposed a method to solve this limitation by encoding the analog input signal with pulse width modulation. This strategy comes at the cost of multiple clock cycles for each encoded input but mitigate *I-V* non-linearity. In this chip, each channel includes one read DAC, and 2 write DACs as input circuits. A digital controller converts a 6-bit input into an *n*-element pulse train of identical Return to Zero (RTZ) pulses where *n* is the input data. The digital output from controller drives a 1-bit DAC, which delivers a pulse train of read-voltage pulses to the crossbar row. An advantage of using RTZ pulses is that the non-idealities introduced at pulse transitions are proportional to the input and show up as a gain error that can be canceled in software. Finally, digital-analog conversion can be also avoided by using analog inputs in their digitized form. Each bit from the analog input number is computed sequentially from the least significant bit to the most significant one. This strategy will increase the number of operations to compute a single VMM but will preserve the analog resolution.

2.3.3. Output circuits

Output signals from a RS-based VMM operation are analog currents that needs to be converted into digital numbers. A straightforward solution is to use ADC and Trans-Impedance Amplifiers (TIA). ADC resolution depends directly on both the conductance resolution of each RS element and the VMM size. For example, 1-bit RS conductance with a vector dimension of 256 (256 lines connecting to one bit-line) requires at least 8-bit of resolution to discriminate all output levels. 5-bit RS memories with the same vector dimension requires 13-bit ADC, which represent in itself a serious design challenge to preserve energy consumption/area efficiency. Employing high resolution ADC in such arrays is one option for distinguishing the analog output levels which requires a careful cost and overhead analysis. Many parameters are used to assess the performance of an ADC such as input impedance, supply rejection, metastability rate, power consumption, die area, signal to noise and distortion ratio (SNDR) and etc [81]. In a typical RS-based VMM engine, the most important metrics to consider for the ADCs are their resolution, their sampling frequency (f_s) and their surface area on the die which affect accuracy, throughput and cost respectively. Figure 7 reveals the main aspects trade-off of the ADC published in the International Solid-State Circuits Conference (ISSCC) from 1997 to 2020. Technology node is the fundamental factor that constrains the area of an ADC (Figure 7(f), whereas a survey of state-of-the-art ADCs [82] reveals that, for smaller technology node and more

diminutive voltage supply headroom, the power consumption is usually bounded by the thermal noise so that one added bit demands quadrupled power rather than only proportional fCV^2. The ADCs with higher resolutions are slower and less power-efficient (Figure 7(c)) while the ADCs with higher sampling frequency have worse energy-efficiency and lower resolution (Figure 7 (b,d)). The achievable performance of the ADC can be predicted by two well-known the figure of merits (FOM) [83-85].

$$FOM_S(\text{dB}) = SNDR + 10\log_{10}\frac{ERBW}{P}$$

where *ERBW* is the bandwidth of the ADC, *P* is the total power dissipation.

$$FOM_W(\text{fJ/CONV. STEP}) = \frac{P}{2^{ENOB} \times \min(f_s, 2ERBW)}$$

where ENOB is the effective number of bits.

In general, the achievable best FOM_S is decreasing along with the increasing of frequency, e.g. doubling f_s or increasing 1-bit resolution postulates quadrupled power consumption (Figure 7(e)). In addition, reducing FOM_W demands an increase of die-area, e.g. 50% power reduction or 1-bit more resolution need 25% more die-area (Figure 7(f)). Overall, the choice of ADC architecture depends on the needs of the application. If each memristor crossbar word-line or bit-line requires one high-resolution ADC (>10-bit), successive approximation register (ADC) or delta-sigma (DSM) ADC can be utilized as SAR ADC and DSM have slightly smaller form factors (Figure 7(a)) and significantly better SNDR. Voltage control oscillator (VCO) based ADC or SAR ADC are more suitable to smaller technology node implementation since they do not rely on high gain/bandwidth amplifiers that limited by intrinsic transistor gain [86]. If the inference operation takes longer than 10ns, low resolution/high-speed flash ADC can be applied via time-multiplexing to minimize die-area since an 8-bit ADC is usually needed for a typical neural network to achieve more than 90% classification accuracy [76,87]. The best possible ADC performance can be estimated based on the system requirement. A decent system-level design can reduce the needed performance of the ADC significantly. The dashed line shown in Figure 7(b) marked the lowest possible ADC power consumption for a given sampling frequency, and the dashed line shown in Figure 7(c) marks the maximum possible ADC SNDR for a given power consumption limit. Therefore, the trade-off among speed, power, and accuracy of ADC can be described by the following equation,

$$P = FOM_{W,\min} \times 2^{ENOB} \times f_s,$$

where $FOM_{W,\min} = 2\times10^{-15}$ for the best state-of-art ADC designed in 28 nm technology. The relationship between the peak SNDR and ENOB is

$$SNDR_{\max}(\text{dB}) = (ENOB \times 6.02 + 1.76).$$

The dashed line in Figure 7(d) labels inevitable trade-off between the peak SNDR and sampling frequency within the current state-of-art ADC

$$SNDR_{\max}(\text{dB}) = 165 - 10\log_{10} ERBW.$$

Figure 7(f) reveals the trade-off between the energy efficiency matrix and area efficiency. The area and energy efficiency improves with shrinking the technology nodes, while they are roughly bounded by the following relationships,

$$Area\ (mm^2) = A - 10^{-4} FOM_w,$$

where *A* is the technology depended factor that equals to 2×10^{-3} for 14nm technology.

The analysis shown above is used for estimating the performance of relatively low-resolution ADC (<14-bit). For higher resolution ADC (>14-bit), adding one more bit means increasing 6 dB SNR, quadrupled less noise power and four-fold larger overall capacitance, as the thermal noise at the input of the ADC equals to KT/C (where K is Boltzmann constant, *T* is Kelvin temperature, and *C* is the capacitance at the input of the ADC). This relation is well defined by Shreier's FOM[83]. Figure 7(e) shows the relationship between the Shreier's FOM and sampling frequency, the maximum achievable FOM at low frequency (<10MHz) is 192 dB, and at higher frequencies (>10MHz), the best achievable FOM equals to

$$FOM_{S_{max}}(\mathrm{dB}) = 192 - (10\log_{10} ERBW/10^7).$$

An alternative sensing approach is to replace the TIA block by a charge-based accumulation circuit. This strategy was used to cope with pulse width modulation encoding that excludes the utilization of TIA[76]. Note that the same approach could be used along with other encoding techniques such as digitization of inputs and pulse amplitude modulation. To maintain precision of the RS-based VMM hardware, the same trade-off in the ADC resolution with crossbar array size is applicable and requires a design optimization in terms of energy consumption and footprints.

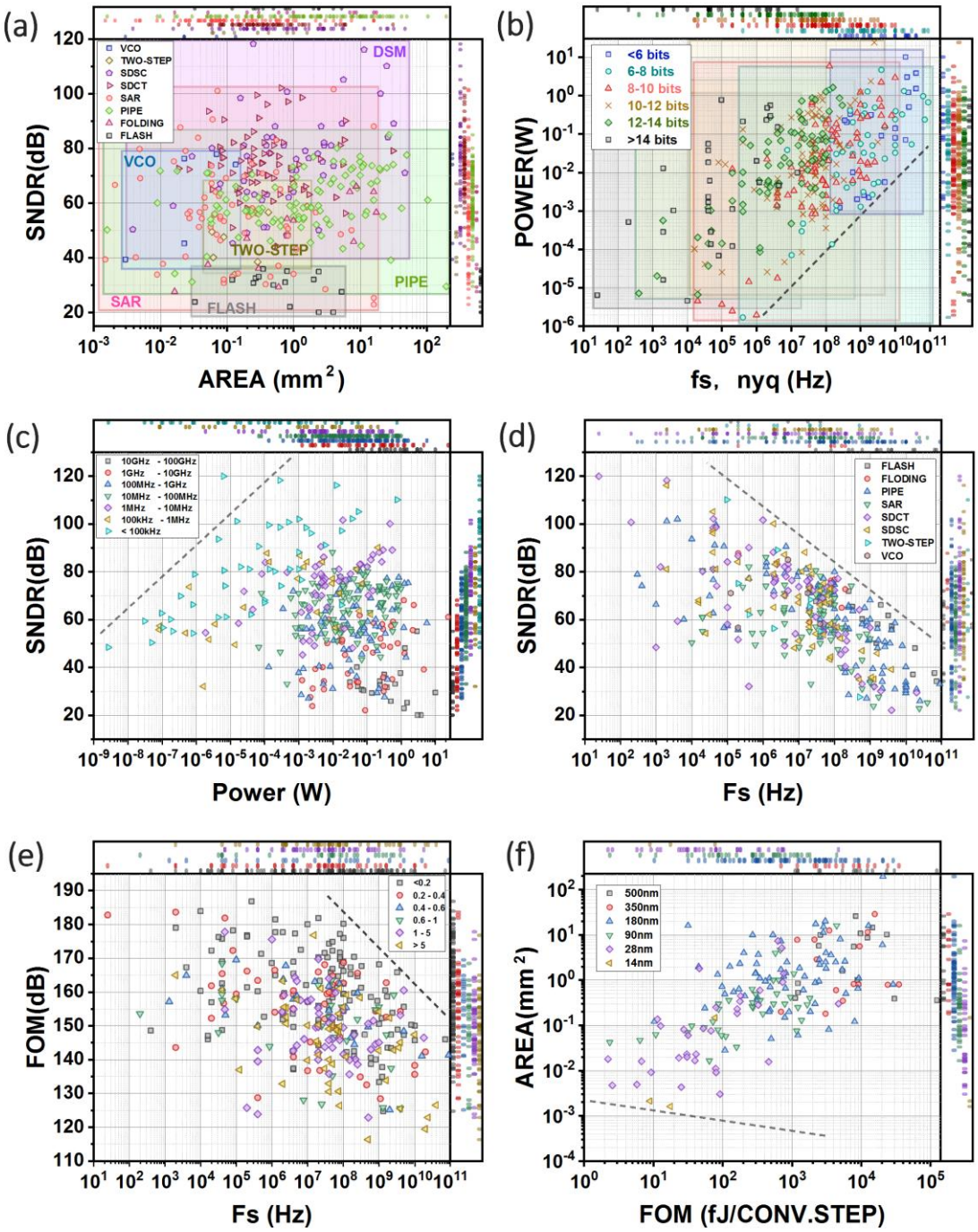

Fig. 7. Main aspects trade-off of the ADC published in the International Solid-State Circuits Conference (ISSCC) from 1997 to 2020. (a) ADC area increases with SNDR and is classified based on different ADC architectures. (b) ADC power consumption increments with the increase of Nyquist sampling frequency and marked based on different resolutions. (c) ADC power consumption versus the SNDR has been displayed for different design operating frequencies. (d) ADC sampling frequency versus the SNDR has been shown for different ADC architectures. (e) ADC FOM versus sampling frequency is displayed and classified based on CMOS technology node (in um). (f) ADC area versus FOM is shown and categorized based on different CMOS technology node.

2.3.4. Recent chips demonstration on integrating CMOS circuits and RS devices

Implementation of VMM hardware using in-memory computing property of RS-based array has become a topic of interest for AI hardware research groups in recent years. Some of these efforts [14,88] have used discrete integrated circuit components connected to the RS array and they did not present a complete integrated system in a single chip. However, there are few fully integrated

CMOS/RS devices chips implemented for VMM-based applications. These fully integrated VMM engines can be categorized into various design choices based on the precision of the selected weight, input and output. However, this categorization can be complemented by considering the classification of these platforms into current-based and time-domain designs. As can be seen in the Figure 8, choices for input, output and weight cell includes binary, ternary, multi-bit and analog. However, selecting a design choice is directly depending on the target application requirements and its functional aspects e.g. accuracy level, speed and etc. There have been several device-, circuit- and system-level concepts proposed to enhance the efficiency and functionality for each of these design choices. As an example, for a binary weight cell design with a circuit level proposition for input and output circuits, a non-volatile intelligent processor (NIP) [89] has been designed by using 4 kb 1T1R binary HfOx-based cells and using 150 nm CMOS technology. This work proposes a non-volatile flip flop circuit by integrating two RS cells into its design for the input and output sensing blocks to avoid high cost DAC and ADC blocks. The output sensing circuit has an adaptive design and can support from 1-bit to 3-bits of resolution. This design improves energy and area efficiency by eliminating the data conversion circuits overhead and turning off the unwanted cells by input-controlled access transistor scheme in 1T1R array. The other physically implemented chip is a binary VMM engine presented in[90] by using 2T2R differential weights with input-controlled access transistor scheme and a pre-charged sense amplifier (PCSA) circuit. This chip was developed for binarized neural network demonstration but consists essentially in a binary dot-product operation. The 2 kb HfOx-based RS devices have been integrated on top of the fourth metal layer in CMOS 130 nm technology node. The PCSA circuit is differential and connected to the both bit-lines of the 2T2R cells in each column. Due to the binarized neural network properties[91], the weights and activation functions are binary and there is no need for multipliers. This design is very efficient for in-memory computing applications where activation functions are implemented by XNOR gates and additions are carried out by popcount gates. This chip is purely digital and it is free from any D/A or A/D conversion which results a high energy and area efficiency performance.

In addition to the mentioned design choices, for ternary weight design, a 1 Mb 1T1R array and its CMOS peripheral circuits were integrated on a single chip in 65 nm CMOS technology node[77]. This implementation proposed new circuit peripherals and architecture level idea to enhance the area and energy efficiency. This platform implements configurable logic operations (XOR, AND and OR) in addition to inference operation. Binary inputs and ternary weights are implementing inference with positive and negative weights located in two separate sub-arrays. Partial MAC results computed from each sub-array are added together to compute a partial MAC. To avoid using costly DAC circuits, this work proposes Dual Word Line Driver (D-WLDR) circuit to apply inputs in both memory and inference modes. These circuits include small digital buffers occupying small area and fitting with the pitch size of the 1T1R cell in the word line. To overcome the issue of area efficiency due to high precision ADC blocks and to enable a highly parallel inference operation, small offset current mode sense amplifier (ML-CSA) and input-aware reference current generator circuit (MIA-RCG) are proposed. MIA-RCG is generating various reference currents in reference arrays to increase the bit-line signal margin between different states for each mode of operation (logic or inference). ML-CSA is minimizing the offset in sense amplifier due to the mismatch of CMOS devices in the bit-line. To further, enhance the readout accuracy and tolerance for small read out margin, Distance Racing Current Mode sense amplifier (DR-CSA) is proposed and shows an improvement in sensing margin by two times in comparison with the mid-point sensing scheme. The platform demonstrates a promising energy efficiency and inference accuracy for various precision values (1-, 2- and 3-bit), but with limited array size (VMM is limited to dimension 12). In the other work[92], a 158kb VMM

engine is designed in 130 nm CMOS technology and it is tried to mitigate the issue of large sensing current in the columns, ADC circuit overhead, problem of voltage drops and transient error of MAC operation in large VMM. A signed weight 2T2R cell has been used in order to reduce the column's sensing current by getting benefit from the differential current. In this work, a quasi-3-bit weight (7-level) is used by positive and negative 1T1R cells which locally cancels their current in the shared column and this should fairly solve both problems of large sensing current and voltage drop impact. This work also presented a low power adjustable resolution ADC circuit (LPAR-ADC) which is reconfigurable from 1-bit to 8-bit precision. The integration and quantization scheme in LPAR-ADC suppressed overshoot and fluctuation of the sensing current which improves the transient error due to the sensing stage. The proposed VMM engine is providing a high energy efficiency of 78.4 TOPS/W when sensing the output by 1-bit precision and high inference accuracy around 94% for MLP of MNIST classification task with 8-bit sensing precision in both ADC stages of the network.

For multi-level weight design choice, [54] proposed a hardware implementation of CNN using 1T1R RS-based VMM engine in 130 nm CMOS technology node. In this hardware, eight 2 kb processing element chips have been integrated on a custom designed PCB to implement a five-layer CNN network. Each of these PE chips, in addition to the RS-based array, includes switching matrix circuits for input and output, 8-bit ADC and shift and add blocks. 4-bit differential pair of 1T1R cells is deployed as weights by tuning the 8-level RS devices. Analog inputs are encoded into 8-bit binary sequential pulses in eight time-intervals and applied via external voltage generator to PE chips. Each PE chips include 4 ADC blocks with 8-bit precision to sense 128×16 RS array. Each ADC block is shared between 4 columns by sample and hold (S/H) circuits for time multiplexing to reduce the overhead cost of the analog to digital conversion. To reduce the latency of inference, each of these 4 columns are connected via a pair of S/H blocks. In first inference step, one S/H block in each pair is sampling the output of its corresponding column. During the next inference step the other S/H block in each pair samples the output while the ADC carries out sensing of the output from all four blocks which sampled in the previous inference cycle (first inference output). This inference scheme reduces the inference latency by pipelining the computation. Hybrid training scheme is utilized to avoid accuracy loss due to the device- and array-level imperfections. This was done by mapping the ex-situ weights on all PE chips in initial steps and subsequently apply multiple runs of in-situ learning on the shared fully connected layer PE chips. This VMM engine design has a very high computational efficiency (1.164 TOPS/mm^2) and energy efficiency (11 TOPS/W) and it enhanced the inference accuracy for MNIST classification task up to 95.57%.

First demonstration of VMM engine with analog weight deploying a passive RS crossbar by the size of 54×108 monolithically integrated with CMOS in 180 nm technology node on a single chip is presented in[76]. In this work, a charge-based inference is targeted to overcome the *I-V* non-linearity of the RS devices. In this context, the analog input is encoded by applying the discrete-time pulse train with the fixed-amplitude into a 6-bit time-domain DAC. The DAC then apply the corresponding 6-bit width modulated input pulse into the array. The bit-line accumulated charges are sensed by an incremental charge-integrating ADC. High resolution hybrid 13-bit ADC circuit is placed in both rows and column to enable bi-directional inference operation and it is comprised of 5-bit first order incremental ADC, 8-bit SAR ADC and additional 1-bit redundancy stage. OpenRISC processor with 64 kB SRAM along with timing generation blocks has been integrated in the chip to initiate different operation modes and control of the DAC and ADC blocks. High resolution input and output circuits and bi-directional inference capability make this platform highly flexible to implement different blend of machine learning applications. However, this flexibility brings cost as the number of ADCs is doubled beside the fact that high-resolution ADC consumes more power and area too.

Each of these VMM engine design choices offers different performance behavior and making a trade-off between accuracy, energy efficiency and area efficiency by considering the application constraints and demands will be vital for the appropriate selection. The detail specification and performance of hardware implemented RS-based VMM engines is presented in Table 1.

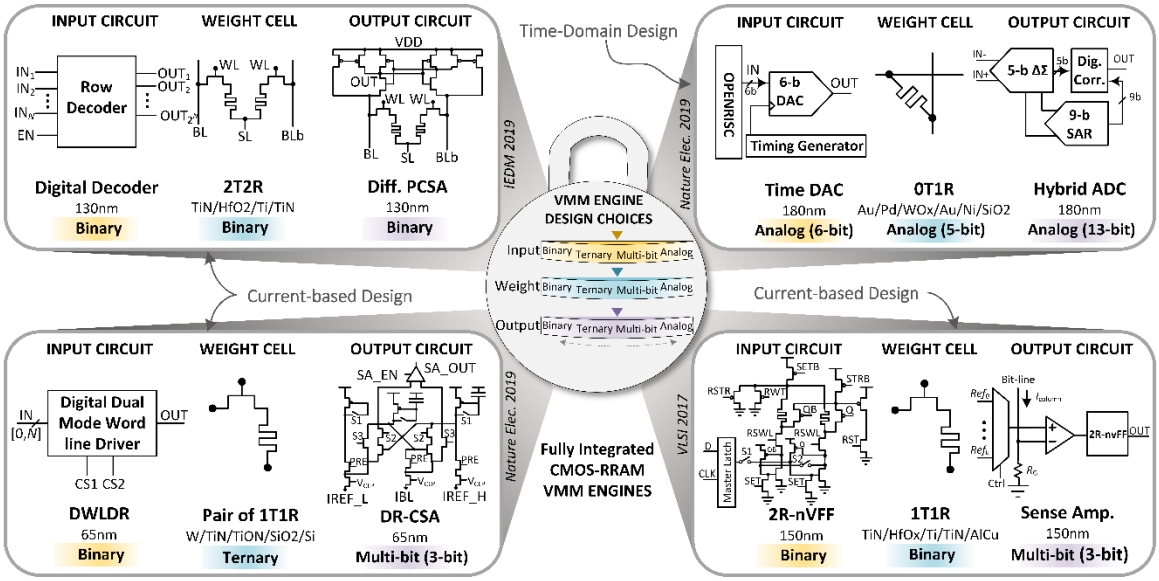

Fig. 8. Design choices for RS-based VMM engines are defined based on the combination of the input, weight cell and output precision targeted for specific applications. Here a combination lock is displayed as VMM design choices which may be unlocked with different combination of the input, weight and output. The input, weight and output choices are binary, ternary, multi-bit and analog input. Here, we have displayed four examples of different design choices from the recent fully integrated CMOS/RS-based chips [76,77,89,90].

Table 1: Comparison of in-memory computing hardware with non-volatile memory blocks by considering capacity of larger than 1 kb.

	Nature Elec. 2017 14	**VLSI 2017** 89	**VLSI 2018** 93	**Nature 2018** 57	**Nature El. 2019** 77	**ISSCC 2019** 94	**ISSCC 2019** 48	**IEDM 2019** 90	**Nature El. 2019** 76	**ISSCC 2020** 92	**Nature 2020** 54
CMOS Technology	2um	150nm	40nm	90nm	65nm	55nm	130nm	130nm	180nm	130nm	130nm
RS Device Type	RRAM	RRAM	RRAM	PCM	RRAM	RRAM	RRAM	RRAM	RRAM	RRAM	RRAM
RS Device Material	Pd/Ta /HfOx /Ta	AlCu/Ti N/Ti/ HfOx/Ti N	XX/TaO 5/TaOx /XX	?	W/TiN/ TiON/Si O2/Si	W/TiN/ TiON/Si O2/Si	TiN/Hf Ox/ Ti/TiN	TiN/Hf Ox/ Ti/TiN	Au/Pd/ WOx/A u	XX/TaOx/H fOx/XX	TiN/TaOx/Hf Ox/TiN
Fully Integrated Chip	NO	YES	YES	NO	YES	YES	YES	YES	YES	YES	Yes
Crossbar Architecture	1T1R	1T1R	1T1R	2PCM + 3T1C	1T1R	1T1R	1T1R	2T2R	0T1R	2T2R	1T1R
# Synapses	8k	4k	4M	524k	1M	1M	18k	1k	6k	158.8k	16k
Weight Resolution	Analog (6-bits)	Binary (1-bits)	Analog (4-bits)?	Analog (>8-bit)	Ternary	Multi-bit (3-bit)	Multi-bit (2.3-bits)	Binary (1-bit)	Analog (6-bits)	Multi-bit (3-bit)	Analog (4-bit)
Input Resolution	Analog	Binary (1-bit)	Binary (1-bit)	Analog * (TD 9-bit)	Binary (1-bit)	Binary (1-bit) & (2-bit)	Binary (1-bit)	Binary (1-bit)	Analog (TD 6-bit)	Binary (1-bit)	Digitized (8-bit)
Output Resolution	Analog(5-8 bits)	Multi-bit (3-bit)	Binary (1-bit)	Analog (8-bit)	Multi-bit (3-bit)	Multi-bit (3-bit)	?	Binary (1-bit)	Analog (13-bit)	Analog (1-bit-8-bit)	Analog (8-bit)
Area (mm^2)	10.9*	3.69	2.71	5.8	6	7.5	11.25	0.2	61.4	21.82	0.0708
Storage Efficiency (Mb/mm^2)	0.0007	0.001	1.47	0.088	0.16	0.133	1.6	0.005	0.00009	0.0072	0.23
Throughput (TOPS)	1.64	0.101	0.66	20	0.019	0.012	0.78 (23 ns Readout)	0.0027	0.057	1.5	0.081
TOPS/mm^2	0.15	0.002	0.24	3.44	0.003	0.0016	0.069	0.013	0.0009	0.071	1.16
Energy Efficiency (TOPS/W)	119.7	0.462	66.5	27.4	16.95	53.17 & 21.9	1.65 (0.47W)	4.2	0.1876 (306 mW)	78.4	11
TOPS/Wmm^2	10.98	0.125	24.5	3.6	2.82	7.08 & 2.92	0.147	20.4	0.003	3.59	156

2.4. SYSTEM LEVEL DEVELOPMENT OF RS-BASED VMM ENGINES.

2.4.1. Leveraging the cost of mixed analog/digital approaches and data trafficking

Performances of VMM engines appears to be strongly affected by the analog-to-digital and digital-to-analog conversion operations, even if the analog MAC operation by itself is very energy efficient. Note that this trade-off between in-memory computing of the MAC operation and overhead circuits cost should evolve favorably by increasing the dimensions of RS-based VMM engines. Indeed, as $N + M$ DACs and ADCs are required to drive a $N \times M$ crossbar array, the energy consumption and the analog/digital interface circuitry per operation should be thus decreased in the case of large-scale

VMM engines. This is to be analyzed in the light of the important challenges that crossbar arrays scaling is facing (see discussion in section 1.3) and represents a vital point for the development of future RS-based VMM engines.

In the previous sections, we pointed out the important trade-off between in-memory computing of the MAC operation with the overhead circuitry required to drive the crossbar array. The proposed analysis considers only the potential improvement in terms of energy and speed offered by computing the MAC operation physically. It doesn't consider energy consumption associated to data trafficking at higher levels, which has been identified in conventional computing platforms (e.g. GPU) as the most expensive operation. Moving data corresponds to both moving parameters of the MAC operation (e.g. matrix components) but also the input and output data (e.g. vectors to be computed / output vectors of the VMM). As can be seen in Figure 10(a-b), in the conventional von Neumann computing systems and Near-Memory Computing (NMC) architectures all input, weigh and output data are moving between the processing unit and memory. However, the traveling distance in NMC systems are significantly smaller than the conventional von Neumann computing architectures. On the other hand, In-memory computing of the MAC operation is proposing to store "permanently" the matrix component into a dedicated non-volatile memory and thus reducing drastically the data movement for these parameters (Figure 10 (c)). Nevertheless, I/O data still needs to be moved and can represent the main bottleneck of the overall system. Note that I/O data can also be used numerous times in the system for specific applications such as convolutional neural network and could be benefited from limited movements (i.e. data re-use). A more detailed analysis of this case needs to be considered for assessing the overall performances of RS-based VMM and system level analysis should address this question.

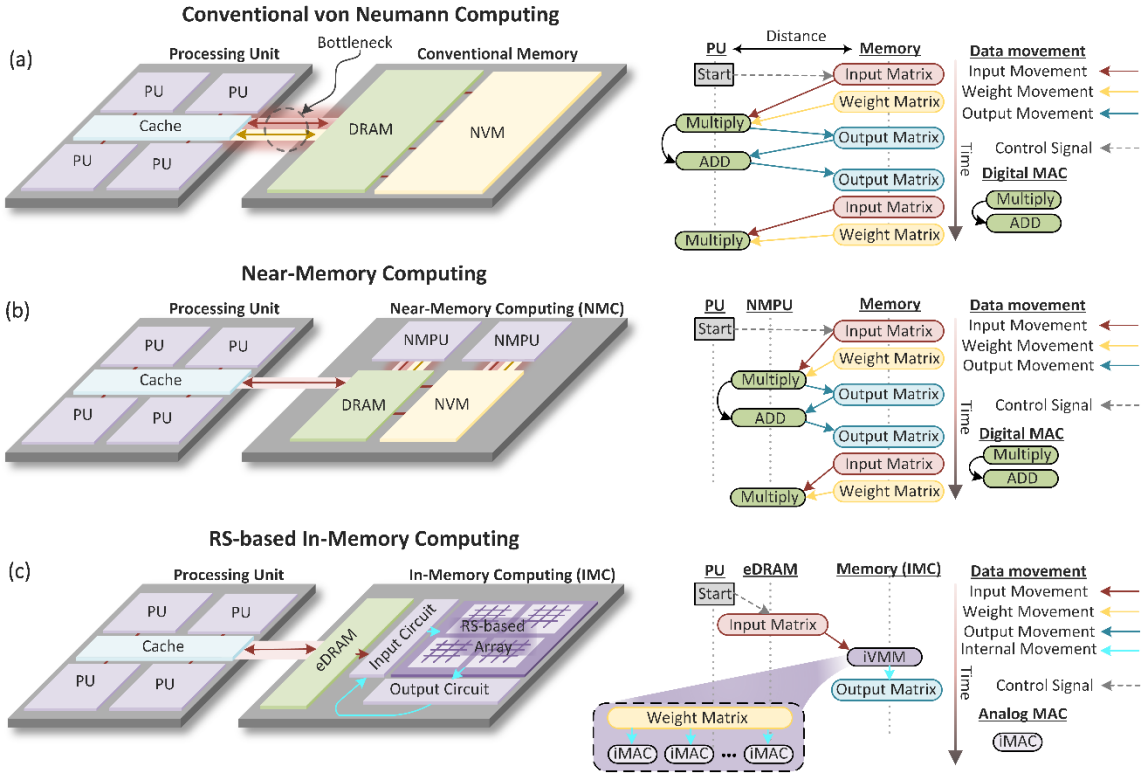

Fig. 10. Three different computing architectures are displayed with their corresponding data movement (input, weight and output) to carry out vector matrix multiplication operation. (a) Conventional von Neumann computing architecture is displayed comprising of processing unit and conventional memory. It

has been shown that a high data movement for both inputs and weights as data needed to be fetched from or stored in the memory at different stages of the operation. Also, the digital MAC increases the computation time as several consecutive digital operation will be needed to perform large VMM. (b) Near-memory computing architecture (NMC) is displayed in this part and, in addition to the main processing unit, near memory processing units (NMPU) have been placed in vicinity of DRAM and NVM blocks. This reduces the data movement cost significantly as the commute distance of data is reduced by placing the processing unit close to the memory. Although in NMC the distance of memory to processing unit is decreased, there is still a significant amount of data commute in between for input, weight and output data. Also, the problem of high computation time due to the digital MAC exists. (c) In-memory computing architectures (IMC) implement computing within the memory. Specifically, in RS-based IMC, the RS-array can implement highly parallel VMM operation in one step and it also stored the weight matrix which will completely omit the weight movement during the operations. The only data movement in the IMC corresponds to input data. In-memory VMM (iVMM) is implemented over RS-based array in fully parallel manner by implementing several parallel in-memory MAC (iMAC) operations.

2.4.2. Current system-level propositions for RS-based VMM engines

In addition to physically implemented RS-based VMM engines, there are promising system-level propositions that are considering more complex ADC optimization and shared circuitry which could be viable for designing very energy efficient APUs. APUs are specialized hardware with a better performance in comparison with CPUs and GPUs to carry out specific tasks and applications. As can be seen in Figure 11(a), RS-based APUs are categorized based on the precision of their weight cells into binarized, ternary, multi-level and analog weight networks. The possibility of implementing wider ranges of applications with high resolution weight networks bring more flexibility in comparison with lower precision peers e.g. binarized and ternarized weight networks. On the other hand, low resolution APUs provides better energy efficiency and lower CMOS circuitry overhead which results in higher storage efficiency.

One of the notable RS-based systems is ISAAC which is a convolutional neural network accelerator[97]. ISAAC consists of tiles which includes eDRAM buffer, pooling unit, adders, and In-situ Multiply Accumulate (IMA) units. Inputs are sent through the eDRAM to IMA units which consist of RRAM crossbars and peripheral circuits (e.g. DAC and ADC) in a h-tree network topology. The dot-product computation of each crossbar is stored in the local sample and hold block. Subsequently, the 8-bit ADCs and shift-and-add circuits are carrying out the digitized outputs computations. This platform applied 16-bit input by digitizing it into 16 cycles of 1-bit pulse generated with 1-bit DACs. Also, 16-bit weights are distributed in 8 columns with each RRAM cell providing 2-bit precision. Further enhancement of ISAAC has been proposed[98], which utilized various ADC optimization techniques such as adaptive ADC scheme and different multiplication methods (e.g. Karatsuba[99] and Strassen's algorithm[100]). This approach reduces the ADC computational overhead and leverage analog resolution of the MAC operation. In addition to these, Newton proposed buffer management techniques and new mapping scheme to overcome the data communication and storage problems, respectively.

The other important system to be noted here is PRIME[101] which is a general platform enabling both memory and computation modes by deploying three RS-based sub-arrays as its memory bank: memory sub-array, FF sub-array and buffer sub-array. FF sub-array is utilized for both storage and computation purposes, memory array is employed only for storage purpose and buffer sub-array is used as the data buffer for FF sub-array. These three sub-arrays have been proposed as an

optimization strategy for data trafficking. In terms of circuit overhead for RS-based VMM operation, PRIME avoids the need for high cost ADC circuits with reconfigurable precision (up to 8-bit) by designing a specific sense amplifier circuit block whose precision is controlled with a counter. Since PRIME has been proposed as a ML-specific platform, rectified linear unit (ReLU) activation function and a block to support max pooling is added after the sense amplifier circuit to provide more efficient properties for applications like CNN.

Alternatively, 3D-aCortex architecture[102] based on 3D NAND flash memories proposes to use time- domain encoding of the information that relax drastically the cost of digital/analog conversions. In this strategy, both input and resulting output are consistently encoded into the pulse width allowing to pipeline multiple VMM operations without converting data back into the digital domain. 3D-aCortex has been presented as a 3D integrated version of 2D-aCortex [103] which is a current-based architecture based on 2D NOR-flash memories and offers more than two orders of magnitude better area efficiency while maintaining same throughput at the cost of low energy efficiency degradation in comparison with its 2D version. However, integrating the partial sums in the output for this time-domain designs requires a large capacitor which is a bottleneck in terms of energy and area efficiency for large sized VMM. To overcome this problem, SIR VMM approach has been proposed in[104] based on the successive integration and rescaling (division) of the input bits. Unlike the previous time-domain encoding techniques, each bit of the digital input is encoded into binary pulses. To reduce the size of the load capacitor, in addition to this successive scheme, the accumulated charges will be divided via charge sharing mechanism. Utilization of SIR approach on the same architecture of 2D-aCortex using 1T1R 4-bit cells provide around 2.5 times higher energy and area efficiency in comparison with conventional VMM methods. Three different design concepts of VMM engines for analog/digital input encoded by amplitude, analog input encoded by pulse duration and digital input encoded by duration is displayed in Figure 11(b-d).

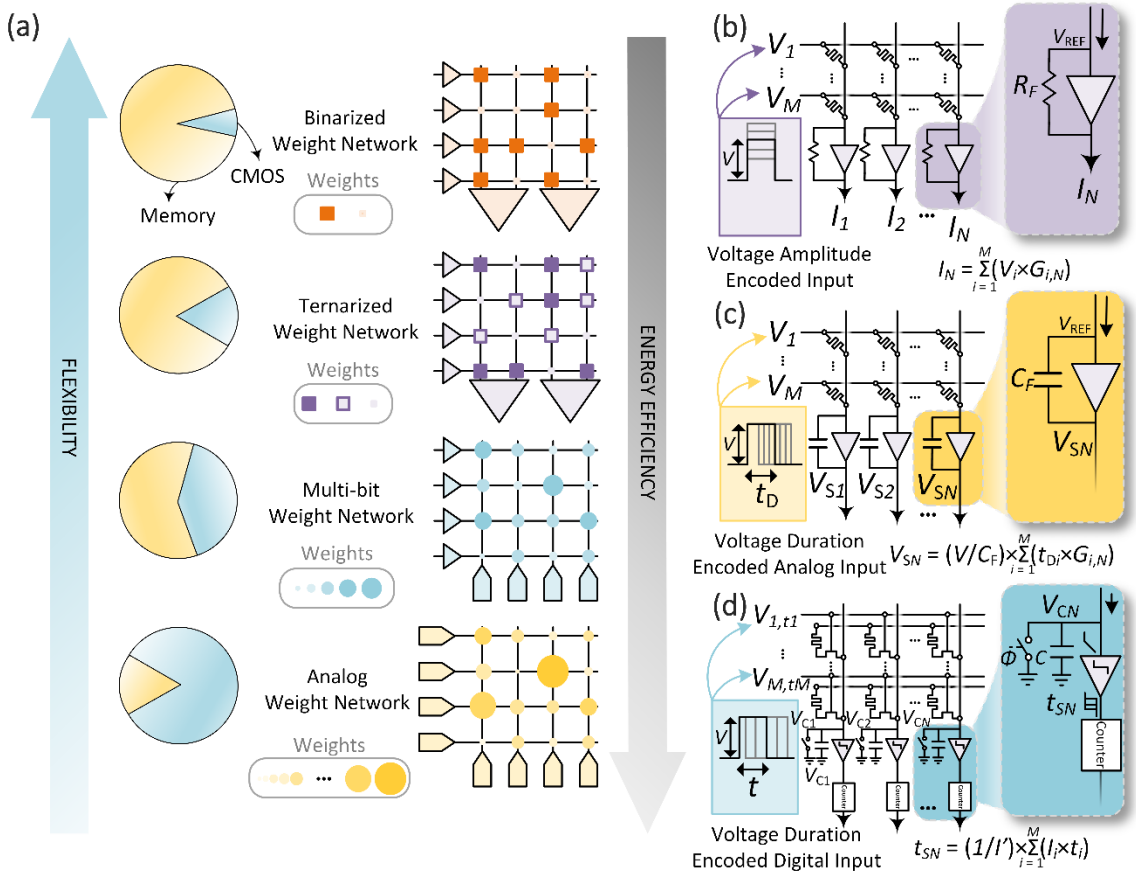

Fig. 11. . (a) Different RS-based APUs are compared in terms of energy efficiency, storage efficiency and flexibility. Each implementation presents a balance between memory functionalities (from binary to analog) and CMOS circuits overhead complexity and cost (b) VMM engine design[88] with 0T1R analog weight network is displayed based on input amplitude encoding and its corresponding sensing circuit with feedback resistor in an op-amp follower block in the bit-line. (c) VMM engine design[95] with 0T1R analog weight network by utilizing input pulse duration encoding scheme and its corresponding sensing circuit for amplitude encoded analog output is displayed. (d) VMM engine design concept[96] for 1T1R weight network by using digital input pulse duration encoding scheme and its corresponding sensing circuit for pulse duration encoded digital output.

---------- BOX2----------
-

Performance Metrics Discussion for ML Accelerators:

Evaluation of the AI accelerators performance for training and inference in ML is a key step in today's competitive race toward building future AI platforms. Performance can be measured for various aspects like Inference Accuracy (IA), Storage Efficiency (SE), Energy Efficiency (EE), and Computational Efficiency (CE). Specific applications will favor some performance metric to another depending on the application constraint (e.g. embedded, high precision computing, low power, …). Specialized hardware developed for ML applications are considering various precision, from 32-bit floating point to binary that makes consistent comparison of IA challenging. As a rule of thumb, conventional ML algorithm can be implemented with limited accuracy of 8-bit integer without compromising too much inference performance. Lower accuracy requires to adapt significantly the algorithms and becomes consequently much more specialized to a specific application. For CE, the important metrics is the throughput, which defines the number of trainings/inferences that can be

carried out by the training/inference engine in a certain amount of time. Conventionally, the numerical computing performance of digital computing systems is measured in Floating Point Operations Per Second (*FLOPS*). However, due to IA inhomogeneity, throughput unit is usually considered as Terra Operations Per Second (*TOPS* or *TOP*/*s*) for ML accelerators. Also, evaluation of the hardware throughput performance accounting for integration efficiency considers *TOPS*/*mm*^{2}. Regarding EE, the number of inference operations is normalized by energy consumption and results in *TOPS*/*W* (*TOP*/*s*/*W* or *TOP*/*J*). Finally, storage efficiency tracks the on-chip memory capacity for weights per unit area and is defined in *MB*/*mm*^{2}.

In addition to *TOPS*, the term *TMACS (Tera Multiply ACcumulates per Second)* is widely used for defining the throughput of the digital neural network (NN) processors which are mostly focused on convolution-centric applications. In the digital APUs inference accelerator as depicted in Figure 9(a), the Multiply Accumulate (MAC) operation consists in successive multiplication and addition operations. This means that, when accelerator manufacturers report the performance of their accelerator in *TMACS*, this value is equal to 2 times the performance in *TOPS*. While in analog VMM engines the MAC is considered as one operation (Figure 9(b)) which is a simple summation of currents over each synaptic device in the bit-line.

In Figure 9(c), the performance comparison of the state of art inference accelerators have been displayed based on throughput and energy efficiency metrics. In this comparison figure, we mostly selected the inference accelerators and tried to include different blend of designs including system solutions: ISAAC[97]NEWTON[98], 2D-aCortex[102], 3D-aCortex[103], SIR[104], PUMA[107], CMOS-based application specific integrated circuits (ASICs): ENVISION[108], AIStorm[109], DNPU[110], UNPU[111], EIE[112], TrueNorth[113], THINKER[114], EdgeTPU[115], TPU[116], Cambricon[117], GOYA[118], QUEST[119], PuDianNao[120], MovidiusX[121], DianNao[122], ShiDianNao[123], DaDianNao[124], RockChip[125], EYERISS V1[105], EYERISS V2[106], and fully integrated CMOS-RRAM VMM chips: UMich[76], Panasonic[93] , Tsinghua chips[54,77,89,92], CNRS[90], IBM PCM[57] by considering their peak performance values.

Despite the common usage of the *TOPS* for ML processing unit evaluation, there is a concern whether this figure is not sufficiently comprehensive for direct evaluation with respect to a given application. For instance, embedded applications do not need necessarily to maximize throughput but would require more drastic limitation in energy consumption. In addition, these reported performance numbers depend on various factors such as network compatibility with the computing platform. For example, different deep NN with the same number of MACs may result in different throughput performance number on the same computing platform.

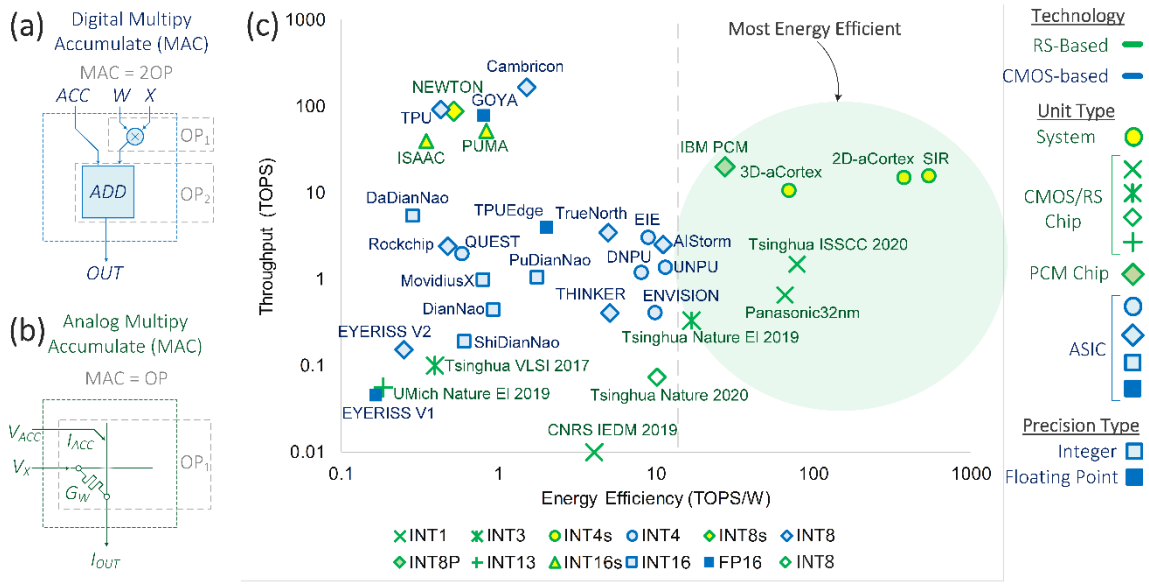

Fig. 9. Implementation of multiply accumulate operation is depicted for both digital and analog domain. Also, the AI accelerator performance comparison is presented. (a) The digital implementation of MAC operation consists of two computational steps, multiplication and addition. Each of these steps are considered one operation (OP). Therefore, digital MAC is two OPs. (b) The analog implementation of MAC is displayed on RS-based array by using Ohm's law and Kirchhoff's law in one computational step. The analog MAC unlike digital MAC is one OP. (c) The inference accelerator performances have been compared in terms of throughput and energy efficiency. The conventional CMOS-based digital ASIC chips, system solutions and RS-based chips are compared by considering the computation precision. Some of these systems or chips are reporting different performance numbers for multiple computation precision while here we demonstrate their performance for one of their reported precisions. Also, this plot may not be a full picture to show these chips and systems performance. As an example, although Eyeriss chips V1 [105] and V2[106] are showing a low energy efficiency below 0.5 TOPS/W in comparison with other systems but they are very low power e.g. Eyriss V1 spends only around 1.67 pJ per MAC operation. This plot shows RS-based systems and chips are the most energy efficient ones. Newton[98], PUMA[107] and ISAAC[97] are also show promising throughput performance in comparison with state of art CMOS-based ASIC chips.

2.5. CONCLUSIONS AND PERSPECTIVES

The competition toward an ideal VMM engine with high performance metrics is an ongoing race between research groups and companies these days. However, lots of factors need to be considered to achieve high reported performance numbers for each of these hardware. In order to reach the reported performance numbers, overcoming the common problems results in reducing the throughput of the deep network inference is primary. Memory access is a limiting factor for achieving high processing speed for the processor as it is going to dominate the computation latency. Increasing the memory bandwidth, reducing the number of memory access in the DNN implementation by scheduling the computation steps, and increasing the arithmetic intensity of the layers which defines the ratio of the computation over the memory access are some of the possible solutions to reduce this effect on the accelerator throughput. To more tighten the gap of the tested throughput with the reported amount, there are some other strategies which needs to be mentioned like maximizing the parallelism to benefit from the full capacity of the hardware resources, reducing the input data transfer time, considering cooling and thermal envelop factor

and heterogonous structure of todays' processors. The approaches described above in this manuscript are examples of generic VMM engines that could be embedded within a digital platform. In order to sustain performances improvement, future hardware deployment based on the basic VMM operation should consider more specialized VMM engine designed for a specific application.

Since RS-based VMMs are analog engines, a clear benefit would be to eliminate Analog/Digital conversions. There are numerous analog applications that could benefit from a local pre-processing of signals based on VMM operation. For instance, RS-based VMM could be embedded into the front-end of sensors networks to compute directly analog signals. Other very demanding applications in terms of VMM operation are ML algorithms. Both synaptic weights and neurons are intrinsically analog elements. By integrating the analog neuron models directly into hybrid CMOS/RS processors, these platforms could maintain ultra-low power consumption and take advantage of purely analog computing. Note that spiking neural networks (e.g. neuromorphic hardware) would benefit from the same scheme since implementing digitally bio-realistic spiking neurons can become very costly while analog approaches seem very efficient. The trade-off here is to favor performances to flexibility since neurones models need to be specified a priori.
From the other side of the spectrum, VMM engine can also be adapted to pure digital operation. For instance, binarized neural networks are machine learning models implemented with simple digital activation function (i.e. neurons), binary input vectors and binary weights. They cannot be used to map all ML algorithms but have demonstrated high performances for tasks that can tolerate binarized data. Their physical implementation with RS-based VMM are highly cost effective and doesn't suffer from limitations such as accuracy and digital/analog conversion. Implementation of the neuron function with CMOS is based on simple XOR majority gates and digital memory in 1T1R configuration for the weights. This approach is somehow reminiscent of biological neural networks that are operating with low resolution synapses and digital action potentials. Including the time encoding strategy used in biological networks into BNNs could lead to an interesting physical implementation of bio-inspired computing. This strategy could potentially reconcile energy efficiency and flexibility of biological computing system that are still the most inspiring objective for future hardware development.

References:

1. Moore, G. E. Cramming more components onto integrated circuits. Proc. IEEE 86, 82–85 (1998).
2. Horowitz, M. 1.1 computing's energy problem (and what we can do about it). In 2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 10–14 (IEEE, 2014).
3. Esmaeilzadeh, H., Blem, E., Amant, R. S., Sankaralingam, K. & Burger, D. Dark silicon and the end of multicore scaling. In 2011 38th Annual international symposium on computer architecture (ISCA), 365–376 (IEEE, 2011).
4. Von Neumann, J. First draft of a report on the edvac. IEEE Annals Hist. Comput. 15, 27–75 (1993).
5. Keckler, S. W., Dally, W. J., Khailany, B., Garland, M. & Glasco, D. Gpus and the future of parallel computing. IEEE micro 31, 7–17 (2011).
6. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. nature 521, 436–444 (2015).
7. Mutlu, O., Ghose, S., Gómez-Luna, J. & Ausavarungnirun, R. Processing data where it makes sense: Enabling in-memory computation. Microprocess. Microsystems 67, 28–41 (2019).
8. Sze, V., Chen, Y.-H., Emer, J., Suleiman, A. & Zhang, Z. Hardware for machine learning: Challenges and opportunities. In 2017 IEEE Custom Integrated Circuits Conference (CICC), 1–8 (IEEE, 2017).
9. Ielmini, D. & Wong, H.-S. P. In-memory computing with resistive switching devices. Nat. Electron. 1, 333–343 (2018).
10. Le Gallo, M. et al. Mixed-precision in-memory computing. Nat. Electron. 1, 246 (2018).
11. Strukov, D., Indiveri, G., Grollier, J. & Fusi, S. Building brain-inspired computing (2019).
12. Sebastian, A., Le Gallo, M., Khaddam-Aljameh, R. & Eleftheriou, E. Memory devices and applications for in-memory computing. Nat. Nanotechnol. 1–16 (2020).

13. Haj-Ali, A., Ben-Hur, R., Wald, N., Ronen, R. & Kvatinsky, S. Imaging: In-memory algorithms for image processing. IEEE Transactions on Circuits Syst. I: Regul. Pap. 65, 4258–4271 (2018).
14. Li, C. et al. Analogue signal and image processing with large memristor crossbars. Nat. Electron. 1, 52 (2018).
15. Liu, S., Wang, Y., Fardad, M. & Varshney, P. K. A memristor-based optimization framework for artificial intelligence applications. IEEE Circuits Syst. Mag. 18, 29–44 (2018).
16. Bojnordi, M. N. & Ipek, E. Memristive boltzmann machine: A hardware accelerator for combinatorial optimization and deep learning. In 2016 IEEE International Symposium on High Performance Computer Architecture (HPCA), 1–13 (IEEE, 2016).
17. Mahmoodi, M., Prezioso, M. & Strukov, D. Versatile stochastic dot product circuits based on nonvolatile memories for high performance neurocomputing and neurooptimization. Nat. communications 10, 1–10 (2019).
18. Cai, F. et al. Harnessing intrinsic noise in memristor hopfield neural networks for combinatorial optimization. arXiv preprint arXiv:1903.11194 (2019).
19. Shin, J. H., Jeong, Y. J., Zidan, M. A., Wang, Q. & Lu, W. D. Hardware acceleration of simulated annealing of spin glass by rram crossbar array. In 2018 IEEE International Electron Devices Meeting (IEDM), (IEEE, 2018).
20. Seo, J.-s. et al. On-chip sparse learning acceleration with cmos and resistive synaptic devices. IEEE Transactions on Nanotechnol. 14, 969–979 (2015).
21. Sheridan, P. M. et al. Sparse coding with memristor networks. Nat. nanotechnology 12, 784 (2017).
22. Kavehei, O. et al. An associative capacitive network based on nanoscale complementary resistive switches for memory-intensive computing. Nanoscale 5, 5119–5128 (2013).
23. Eryilmaz, S. B. et al. Brain-like associative learning using a nanoscale non-volatile phase change synaptic device array. Front. neuroscience 8, 205 (2014).
24. Hu, S. et al. Associative memory realized by a reconfigurable memristive hopfield neural network. Nat. communications 6, 1–8 (2015).
25. Wu, T. F. et al. Brain-inspired computing exploiting carbon nanotube fets and resistive ram: Hyperdimensional computing case study. In 2018 IEEE International Solid-State Circuits Conference-(ISSCC), 492–494 (IEEE, 2018).
26. Rahimi, A. et al. High-dimensional computing as a nanoscalable paradigm. IEEE Transactions on Circuits Syst. I: Regul. Pap. 64, 2508–2521 (2017).
27. Eleftheriou, E. et al. Deep learning acceleration based on in-memory computing. IBM J. Res. Dev. 63, 7–1 (2019).
28. Song, L., Qian, X., Li, H. & Chen, Y. Pipelayer: A pipelined reram-based accelerator for deep learning. In 2017 IEEE International Symposium on High Performance Computer Architecture (HPCA), 541–552 (IEEE, 2017).
29. Sebastian, A. et al. Computational memory-based inference and training of deep neural networks. In 2019 Symposium on VLSI Technology, T168–T169 (IEEE, 2019).
30. Gokmen, T. & Vlasov, Y. Acceleration of deep neural network training with resistive cross-point devices: Design considerations. Front. neuroscience 10, 333 (2016).
31. Mahmoodi, M. et al. Ultra-low power physical unclonable function with nonlinear fixed-resistance crossbar circuits. In 2019 IEEE International Electron Devices Meeting (IEDM), 30–1 (IEEE, 2019).
32. Jiang, H. et al. A provable key destruction scheme based on memristive crossbar arrays. Nat. Electron. 1, 548–554 (2018).
33. Nili, H. et al. Hardware-intrinsic security primitives enabled by analogue state and nonlinear conductance variations in integrated memristors. Nat. Electron. 1, 197–202 (2018).
34. Gao, L., Chen, P.-Y., Liu, R. & Yu, S. Physical unclonable function exploiting sneak paths in resistive cross-point array. IEEE Transactions on Electron Devices 63, 3109–3115 (2016).
35. Choi, S., Sheridan, P. & Lu, W. D. Data clustering using memristor networks. Sci. reports 5, 10492 (2015).
36. Choi, S., Shin, J. H., Lee, J., Sheridan, P. & Lu, W. D. Experimental demonstration of feature extraction and dimensionality reduction using memristor networks. Nano letters 17, 3113–3118 (2017).
37. Rahimi Azghadi, M. et al. Cmos and memristive hardware for neuromorphic computing. Adv. Intell. Syst. (2020).
38. Yan, B. et al. Rram-based spiking nonvolatile computing-in-memory processing engine with precision configurable in situ nonlinear activation. In 2019 Symposium on VLSI Technology, T86–T87 (IEEE, 2019).
39. Serb, A. et al. Unsupervised learning in probabilistic neural networks with multi-state metal-oxide memristive synapses. Nat. communications 7, 1–9 (2016).
40. Gupta, I. et al. Real-time encoding and compression of neuronal spikes by metal-oxide memristors. Nat. communications 7, 1–9 (2016).
41. Prezioso, M., Bayat, F. M., Hoskins, B., Likharev, K. & Strukov, D. Self-adaptive spike-time-dependent plasticity of metal-oxide memristors. Sci. reports 6, 1–6 (2016).
42. Sun, Z. et al. Solving matrix equations in one step with cross-point resistive arrays. Proc. Natl. Acad. Sci. 116, 4123–4128 (2019).
43. Zidan, M. A. et al. A general memristor-based partial differential equation solver. Nat. Electron. 1, 411–420 (2018).
44. Moon, J. et al. Temporal data classification and forecasting using a memristor-based reservoir computing system. Nat. Electron. 2, 480–487 (2019).

45. Du, C. et al. Reservoir computing using dynamic memristors for temporal information processing. Nat. communications 8, 2204 (2017).
46. Midya, R. et al. Reservoir computing using diffusive memristors. Adv. Intell. Syst. 1, 1900084 (2019).
47. Chen, Y. Reram: History, status, and future. IEEE Transactions on Electron Devices 67, 1420–1433 (2020).
48. Wu, T. F. et al. 14.3 a 43pj/cycle non-volatile microcontroller with 4.7 ms shutdown/wake-up integrating
2.3-bit/cell resistive ram and resilience techniques. In 2019 IEEE International Solid-State Circuits Conference-(ISSCC), 226–228 (IEEE, 2019).
49. Chua, L. Memristor-the missing circuit element. IEEE Transactions on circuit theory 18, 507–519 (1971).
50. Strukov, D. B., Snider, G. S., Stewart, D. R. & Williams, R. S. The missing memristor found. nature 453, 80–83 (2008).
51. Xia, Q. & Yang, J. J. Memristive crossbar arrays for brain-inspired computing. Nat. materials 18, 309–323 (2019).
52. Alibart, F., Gao, L., Hoskins, B. D. & Strukov, D. B. High precision tuning of state for memristive devices by adaptable variation-tolerant algorithm. Nanotechnology 23, 075201 (2012).
53. Kim, H., Nili, H., Mahmoodi, M. & Strukov, D. 4k-memristor analog-grade passive crossbar circuit. ArXiv preprint arXiv:1906.12045 (2019).
54. Yao, P. et al. Fully hardware-implemented memristor convolutional neural network. Nature 577, 641–646 (2020).
55. Yao, P. et al. Face classification using electronic synapses. Nat. communications 8, 1–8 (2017).
56. Li, C. et al. Efficient and self-adaptive in-situ learning in multilayer memristor neural networks. Nat. communications 9, 1–8 (2018).
57. Ambrogio, S. et al. Equivalent-accuracy accelerated neural-network training using analogue memory. Nature 558, 60–67 (2018).
58. Sassine, G. et al. Interfacial versus filamentary resistive switching in tio2 and hfo2 devices. J. Vac. Sci. & Technol. B, Nanotechnol. Microelectron. Materials, Process. Meas. Phenom. 34, 012202 (2016).
59. Govoreanu, B. et al. 10 10nm 2 hf/hfo x crossbar resistive ram with excellent performance, reliability and low-energy operation. In 2011 International Electron Devices Meeting, 31–6 (IEEE, 2011).
60. Pi, S. et al. Memristor crossbar arrays with 6-nm half-pitch and 2-nm critical dimension. Nat. nanotechnology 14, 35–39 (2019).
61. Lin, P. et al. Three-dimensional memristor circuits as complex neural networks. Nat. Electron. 3, 225–232 (2020).
62. Adam, G. C. et al. 3-d memristor crossbars for analog and neuromorphic computing applications. IEEE Transactions on Electron Devices 64, 312–318 (2016).
63. Burr, G. W. et al. Access devices for 3d crosspoint memory. J. Vac. Sci. & Technol. B, Nanotechnol. Microelectron. Materials, Process. Meas. Phenom. 32, 040802 (2014).
64. Wang, C. et al. Cross-point resistive memory: Nonideal properties and solutions. ACM Transactions on Des. Autom. Electron. Syst. (TODAES) 24, 1–37 (2019).
65. Adam, G. C., Khiat, A. & Prodromakis, T. Challenges hindering memristive neuromorphic hardware from going mainstream. Nat. communications 9, 1–4 (2018).
66. Sung, C., Hwang, H. & Yoo, I. K. Perspective: A review on memristive hardware for neuromorphic computation. J. Appl. Phys. 124, 151903 (2018).
67. Alibart, F., Zamanidoost, E. & Strukov, D. B. Pattern classification by memristive crossbar circuits using ex situ and in situ training. Nat. communications 4, 1–7 (2013).
68. Pan, W.-Q. et al. Strategies to improve the accuracy of memristor-based convolutional neural networks. IEEE Transactions on Electron Devices 67, 895–901 (2020).
69. Chen, P.-Y. et al. Mitigating effects of non-ideal synaptic device characteristics for on-chip learning. In 2015 IEEE/ACM International Conference on Computer-Aided Design (ICCAD), 194–199 (IEEE, 2015).
70. Hu, M. et al. Memristor-based analog computation and neural network classification with a dot product engine. Adv. Mater. 30, 1705914 (2018).
71. Lee, H. et al. Evidence and solution of over-reset problem for hfo x based resistive memory with sub-ns switching speed and high endurance. In 2010 International Electron Devices Meeting, 19–7 (IEEE, 2010).
72. Xia, L., Liu, M., Ning, X., Chakrabarty, K. & Wang, Y. Fault-tolerant training enabled by on-line fault detection for rram-based neural computing systems. IEEE Transactions on Comput. Des. Integr. Circuits Syst. 38, 1611–1624 (2018).
73. Cai, F. et al. A fully integrated reprogrammable memristor–cmos system for efficient multiply–accumulate operations. Nat. Electron. 2, 290–299 (2019).
74. Chen, W.-H. et al. Cmos-integrated memristive non-volatile computing-in-memory for ai edge processors. Nat. Electron. 2, 420–428 (2019).
75. Dozortsev, A., Goldshtein, I. & Kvatinsky, S. Analysis of the row grounding technique in a memristor-based crossbar array. Int. J. Circuit Theory Appl. 46, 122–137 (2018).
76. Carusone, T. C., Johns, D. A. & Martin, K. W. Analog integrated circuit design [m]. john wiley&sons (2011).
77. Hastings, A. The an of analog layout (Prentice hall New Jersey, 2001).
78. Ohnhäuser, F. Analog-digital converters for industrial applications including an introduction to digital-analog converters (Springer, 2015).
79. Murmann, B. ADC Performance Survey 1997-2020 (2020).

80. Schreier, R., Temes, G. C. et al. Understanding delta-sigma data converters, vol. 74 (IEEE press Piscataway, NJ, 2005).
81. Walden, R. H. Analog-to-digital converter survey and analysis. IEEE J. on selected areas communications 17, 539–550 (1999).
82. Harpe, P., Gao, H., van Dommele, R., Cantatore, E. & van Roermund, A. H. A 0.20 text{mm}ˆ23 nw signal acquisition ic for miniature sensor nodes in 65 nm cmos. IEEE J. Solid-State Circuits 51, 240–248 (2015).
83. Rabuske, T. & Fernandes, J. Charge-Sharing SAR ADCs for Low-Voltage Low-Power Applications (Springer, 2017).
84. Hashemi, S., Anthony, N., Tann, H., Bahar, R. I. & Reda, S. Understanding the impact of precision quantization on the accuracy and energy of neural networks. In Design, Automation & Test in Europe Conference & Exhibition (DATE), 2017, 1474–1479 (IEEE, 2017).
85. Bayat, F. M. et al. Implementation of multilayer perceptron network with highly uniform passive memristive crossbar circuits. Nat. communications 9, 1–7 (2018).
86. Su, F. et al. A 462gops/j rram-based nonvolatile intelligent processor for energy harvesting ioe system featuring nonvolatile logics and processing-in-memory. In 2017 Symposium on VLSI Technology, T260–T261 (IEEE, 2017).
87. Bocquet, M. et al. In-memory and error-immune differential rram implementation of binarized deep neural networks. In 2018 IEEE International Electron Devices Meeting (IEDM), 20–6 (IEEE, 2018).
88. Hubara, I., Courbariaux, M., Soudry, D., El-Yaniv, R. & Bengio, Y. Binarized neural networks. In Advances in neural information processing systems, 4107–4115 (2016).
89. Liu, Q. et al. 33.2 a fully integrated analog reram based 78.4 tops/w compute-in-memory chip with fully parallel mac computing. In 2020 IEEE International Solid-State Circuits Conference-(ISSCC), 500–502 (IEEE, 2020).
90. Mochida, R. et al. A 4m synapses integrated analog reram based 66.5 tops/w neural-network processor with cell current controlled writing and flexible network architecture. In 2018 IEEE Symposium on VLSI Technology, 175–176 (IEEE, 2018).
91. Xue, C.-X. et al. 24.1 a 1mb multibit reram computing-in-memory macro with 14.6 ns parallel mac computing time for cnn based ai edge processors. In 2019 IEEE International Solid-State Circuits Conference-(ISSCC), 388–390 (IEEE, 2019).
92. Chen, Y.-H., Krishna, T., Emer, J. S. & Sze, V. Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks. IEEE journal solid-state circuits 52, 127–138 (2016).
93. Chen, Y.-H., Yang, T.-J., Emer, J. & Sze, V. Eyeriss v2: A flexible accelerator for emerging deep neural networks on mobile devices. IEEE J. on Emerg. Sel. Top. Circuits Syst. 9, 292–308 (2019).
94. Nag, A. et al. Newton: Gravitating towards the physical limits of crossbar acceleration. IEEE Micro 38, 41–49 (2018).
95. Ankit, A. et al. Puma: A programmable ultra-efficient memristor-based accelerator for machine learning inference. In Proceedings of the Twenty-Fourth International Conference on Architectural Support for Programming Languages and Operating Systems, 715–731 (2019).
96. Shafiee, A. et al. Isaac: A convolutional neural network accelerator with in-situ analog arithmetic in crossbars. ACM SIGARCH Comput. Archit. News 44, 14–26 (2016).
97. Bavandpour, M., Mahmoodi, M. R. & Strukov, D. acortex: An energy-efficient multi-purpose mixed-signal inference accelerator. accepted, IEEE J. Explor. Solid-State Comput. Devices Circuits (2020).
98. Bavandpour, M., Sahay, S., Mahmoodi, M. R. & Strukov, D. B. 3d-acortex: An ultra-compact energy-efficient neurocomputing platform based on commercial 3d-nand flash memories. arXiv preprint arXiv:1908.02472 (2019).
99. Bavandpour, M., Sahay, S., Mahmoodi, M. R. & Strukov, D. Efficient mixed-signal neurocomputing via successive integration and rescaling. IEEE Transactions on Very Large Scale Integration (VLSI) Syst. (2019).
100. Moons, B., Uytterhoeven, R., Dehaene, W. & Verhelst, M. 14.5 envision: A 0.26-to-10tops/w subword-parallel dynamic-voltage-accuracy-frequency-scalable convolutional neural network processor in 28nm fdsoi. In 2017 IEEE International Solid-State Circuits Conference (ISSCC), 246–247 (IEEE, 2017).
101. Merritt, R. Startup Accelerates AI at the Sensor (2019).
102. Shin, D., Lee, J., Lee, J. & Yoo, H.-J. 14.2 dnpu: An 8.1 tops/w reconfigurable cnn-rnn processor for general-purpose deep neural networks. In 2017 IEEE International Solid-State Circuits Conference (ISSCC), 240–241 (IEEE, 2017).
103. Lee, J. et al. Unpu: An energy-efficient deep neural network accelerator with fully variable weight bit precision. IEEE J. Solid-State Circuits 54, 173–185 (2018).
104. Han, S. et al. Eie: efficient inference engine on compressed deep neural network. ACM SIGARCH Comput. Archit. News 44, 243–254 (2016).
105. DeBole, M. V. et al. Truenorth: Accelerating from zero to 64 million neurons in 10 years. Computer 52, 20–29 (2019).
106. Yin, S. et al. A high energy efficient reconfigurable hybrid neural network processor for deep learning applications. IEEE J. Solid-State Circuits 53, 968–982 (2017).
107. Google. Edge TPU: Google's purpose-built ASIC designed to run inference at the edge (2019).
108. Jouppi, N. P. et al. In-datacenter performance analysis of a tensor processing unit. In Proceedings of the 44th Annual International Symposium on Computer Architecture, 1–12 (2017).
109. Cutress, I. Cambricon, maker of hauwei's kirin npu ip, build a big ai chip and pcie card (2018).
110. Armasu, L. Move Over GPUs: Startup's Chip Claims to Do Deep Learning Inference Better (2018).

111. Ueyoshi, K. et al. Quest: A 7.49 tops multi-purpose log-quantized dnn inference engine stacked on 96mb 3d sram using inductive-coupling technology in 40nm cmos. In 2018 IEEE International Solid-State Circuits Conference-(ISSCC), 216–218 (IEEE, 2018).
112. Liu, D. et al. Pudiannao: A polyvalent machine learning accelerator. ACM SIGARCH Comput. Archit. News 43, 369–381 (2015).
113. Hruska, J. New Movidius Myriad X VPU Packs a Custom Neural Compute Engine (2017).
114. Chen, T. et al. Diannao: A small-footprint high-throughput accelerator for ubiquitous machine-learning. ACM SIGARCH Comput. Archit. News 42, 269–284 (2014).
115. Du, Z. et al. Shidiannao: Shifting vision processing closer to the sensor. In Proceedings of the 42nd Annual International Symposium on Computer Architecture, 92–104 (2015).
116. Chen, Y. et al. Dadiannao: A machine-learning supercomputer. In 2014 47th Annual IEEE/ACM International Symposium on Microarchitecture, 609–622 (IEEE, 2014).
117. Rockchip. Rockchip Released Its First AI Processor RK3399Pro NPU Performance up to 2.4TOPs (2018).
118. Karatsuba, A. A. & Ofman, Y. P. Multiplication of many-digital numbers by automatic computers. In Doklady Akademii Nauk, vol. 145, 293–294 (Russian Academy of Sciences, 1962).
119. Huss-Lederman, S., Jacobson, E. M., Johnson, J. R., Tsao, A. & Turnbull, T. Strassen's algorithm for matrix multiplication: Modeling, analysis, and implementation. In Proceedings of Supercomputing, vol. 96, 9–6 (Citeseer, 1996).
120. Chi, P. et al. Prime: A novel processing-in-memory architecture for neural network computation in reram-based main memory. ACM SIGARCH Comput. Archit. News 44, 27–39 (2016).
121. Marinella, M. J. et al. Multiscale co-design analysis of energy, latency, area, and accuracy of a reram analog neural training accelerator. IEEE J. on Emerg. Sel. Top. Circuits Syst. 8, 86–101 (2018).
122. Bavandpour, M., Mahmoodi, M. R. & Strukov, D. B. Energy-efficient time-domain vector-by-matrix multiplier for neurocomputing and beyond. IEEE Transactions on Circuits Syst. II: Express Briefs 66, 1512–1516 (2019).

3. CHAPTER 3

Filamentary switching: Synaptic plasticity through device volatility

3.1. INTRODUCTION

Massive amounts of heterogeneous data are generated each day in our society. In this context, computing systems face important challenges in providing suitable solutions for information processing. Saturation of conventional computer performances due to material issues (i.e., clock frequency and energy limitations) and more fundamental constraints inherent in the Von Neumann bottleneck have forced researchers to investigate new computing paradigms that will allow for more powerful systems. The bio-inspired approach (or, more precisely, neuromorphic engineering) is a promising direction for such an objective. Recent breakthroughs at the system [ref1], circuit [ref2], and device levels [ref3] are very encouraging indicators for the development of computing systems that can replicate the brain's performances in tasks such as recognition, mining, and synthesis [ref4]. To achieve such an ambitious goal, research efforts are needed for understanding the computing principles of biological systems, elucidating how spike-coding information is computed and stored in neuron and synapse assemblies, and exploring neuromorphic approaches that define hardware functionalities, performances, and integration requirements. Emerging nanotechnologies could play a major role in this context, by offering devices with attractive bio-inspired functionalities and associated performances that would ensure the future development of neuromorphic hardware.

Some studies have investigated the possibility of implementing neurons in nanoscale devices [ref5,6]. Most of these efforts have been devoted to the realization of synaptic elements with emerging memory devices, such as RRAM technologies, with the goals of matching the critical integration density of the synaptic connections [ref7] and replicating the synaptic plasticity mechanisms that correspond to the modification of synaptic conductance during learning and computing. Indeed, modification of the synaptic weight as a function of neuronal activity (i.e., spiking activity) is widely recognized as a key mechanism for the processing and storage of information in neural networks.

Plasticity mechanisms are commonly categorized as short- and long-term plasticity (STP and LTP, respectively). STP corresponds to a neuronally induced synaptic weight modification that tends to relax toward a resting state, thereby providing activity-dependent signal processing. In LTP, the synaptic weight modification can last for days to months. Thus, LTP provides the information storage capability to the network. Spike timing-dependent plasticity is a variation of Hebb's rule [ref8,9] that has attracted a lot of attention. Although not involved in all mechanisms of learning, spike timing-dependent plasticity has been demonstrated in various nanoscale memory or memristive devices [ref10-17]. Other important expressions of plasticity that have been displayed in memristive systems include STP [ref18,19], demonstrated based on the volatile memory effect, and the STP to LTP transition [ref20-23], displayed in filamentary memory devices in which electrical conductivity is modulated by growth of a conductive filament. Conductive filament growth is induced by the accumulation of electrical stress and leads to an increase in device conductivity. By analogy to long-term memorization processes, which involve the accumulation of short-term effects, and to the idea of reinforcement learning [ref24], conductive filament growth has been directly correlated with increased filament stability, corresponding to long-term storage of the conductive state. In these different works, while the strong analogy between biological synapses and nanoscale filamentary memory devices is evidenced,

transition between STP and LTP is intrinsic to the material system considered (i.e. ionic species, ionic conductor) and cannot be controlled and tuned during operation.

In this paper, we demonstrate that more complex plastic behaviors can emerge from nanoscale memristive devices, thus allowing a greater number of features to be embedded in a single component and potentially permitting more complex computing systems. By considering more complex filament shapes, such as dendritic metallic paths of different branch densities and widths, we show that the volatile/nonvolatile regime can be tuned independently, leading to an independent control of STP and LTP.

Based on the observation of metallic filaments in macroscale electrochemical metallization (ECM) cells, we investigated the growth and stability properties of dendritic filaments. The results were used as a basis for the development of nanoscale solid-state synapses that display independent control of STP and LTP processes via spiking excitation and past history modification. When this behavior was interpreted from the framework of the phenomenological modeling developed for synaptic plasticity, the results revealed a strong analogy between our solid-state device and biological synapses. The additional functionality of independent control of STP and LTP could lead to new learning and computing strategies for neuromorphic engineering and artificial neural networks.

3.2. Results and Discussion

3.2.1. Ag2S filamentary switching

The basic structure of the synaptic device (Figure 1a) corresponds to a conventional ECM cell, as described by Waser [ref25]. Inert and reactive Pt and Ag electrodes, respectively, are separated by a Ag_2S ionic conductor material (60 nm), which ensures the migration of oxidized Ag^+ ions between the electrodes. A positive bias (with a grounded Pt electrode) induces the oxidation of Ag into Ag^+ ions at the Ag electrode, the migration of ions from the Ag anode to the Pt cathode, and the reduction of Ag^+ ions into Ag filaments across the insulating Ag_2S, thereby turning the device from an insulating OFF state to a conductive ON state (SET transition). A negative bias induces the oxidation of Ag from the filament into Ag^+ ions and reduction at the Ag electrode, leading to a disruption of the conductive path that turns the device OFF (RESET transition). To gain insight into the filament shape and growth mechanism, we performed optical microscopic imaging during the current-voltage (I-V) measurement on millimeter-scale devices with a square-shaped Pt electrode on top of a Ag/Ag_2S substrate (Figure 1b).

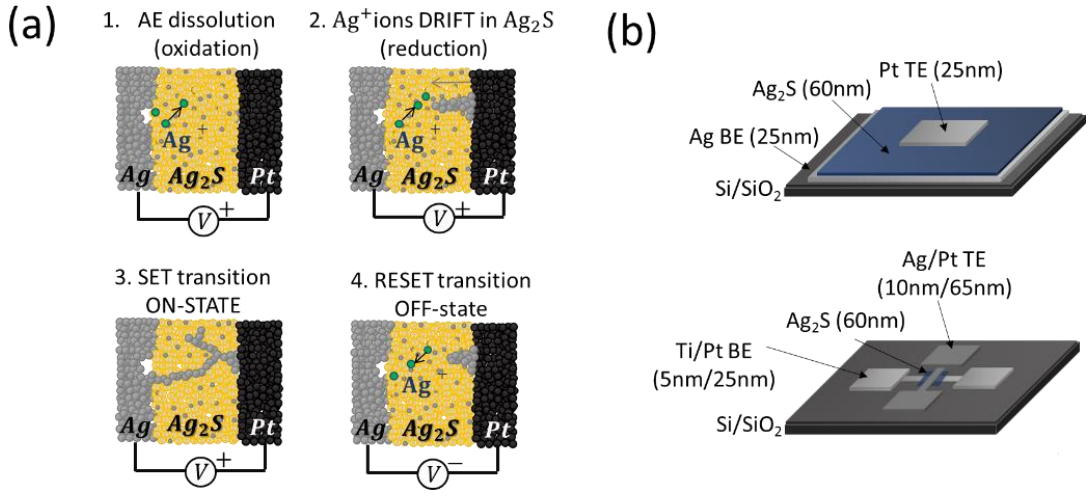

Figure 1: Filamentary switching. (a) Basic switching mechanism of ECM cells. (b) Device configuration at the millimeter scale (top: 0.1 mm × 0.1 mm active area) and nanometer scale (bottom: 200 nm × 200 nm cross-point active area).

Consistent with the switching scenario described above, a positive bias induced the formation of Ag dendritic filaments from the Pt cathode toward the Ag anode (SET transition, Figure 2a, snapshot 1 to 3).

Application of a negative bias induced a partial destruction of the conducting paths, with remaining filament traces corresponding to preferential paths for subsequent switching (RESET transition, Figure 2a, snapshot 4). After an identical positive SET transition, an intermediate situation was observed, in which the device was kept grounded for 5 minutes with a slow dissolution of the metallic dendrites (Figure 2b, snapshot 4*). Such filament relaxation can be attributed to the Ag^+ ion diffusion in the Ag_2S ionic conductor and to the reverse oxidation-reduction process of the Ag filaments [ref26].

A second analysis of the filament formation was realized by varying the compliance current (I_C) during the SET process. This approach is commonly used in ECM cells to tune the conductance of the ON state and to limit the formation of filaments [ref 10.1109/TED.2009.2016019]. If tuning the conductance by limiting the growth of a single filament is considered straightforward (i.e., because the filament diameter corresponds directly to the conductance state), then a more complex picture was obtained for ECM cells that had complex dendritic filament morphologies. Increasing the density or width of the dendritic branch can correspond to an increase of conductance. Due to the resolution of the optical microscope, it was not possible to obtain an accurate assessment of filament diameter. However, we effectively measured a larger filament expansion and dendritic tree density with a larger I_C (Figure 2c). This observation indicates a direct correlation between I_C and the fractal geometry of the dendritic filaments (see Supporting Information, Figure S1). Again, after RESET, the remaining filament traces corresponded to preferential paths for subsequent switching.

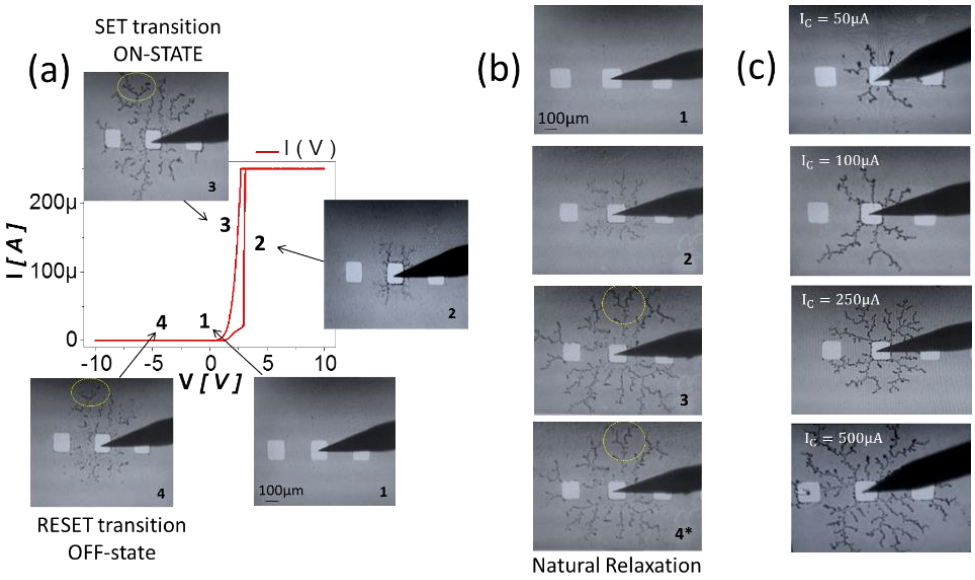

Figure 2: Millimeter-scale ECM cell configuration. (a) I-V characteristics and associated optical microscope imaging (0.1 mm × 0.1 mm) of filament growth. (b) Natural relaxation of the filament. After a positive SET transition (1–3), the device was kept grounded for 5 minutes (4*). (c) Relationship between I_C and dendritic expansion/shape.

Using the previous analysis as a guideline for describing nanoscale filament stability, we implemented the same structure in nanoscale devices consisting of Ag/Pt cross-points with a 200 nm × 200 nm active area separated by Ag_2S (Figure 1b). This device configuration offers the potential for cross-bar integration (cross-point of metallic wires) and for the realization of dense synaptic arrays. Due to the high mobility of the Ag^+ ions in the Ag_2S ionic conductor, the device was operated at low voltages, close to the biological electrical potential recorded in neuronal cells during spiking (200 mV vs. 80 mV).

As expected, controlling the I_C value during SET transition limited the filament growth and tuned the ON conductance state. ON states at I_C values of 100 nA to 50 µA were strongly volatile, whereas ON states at I_C values above 50 µA were stable, with RESET transition observed at a negative bias (Figure 3a). A linear I-V relationship, defining the ON conductance state G_{ON}, was obtained in all ON states, indicating that the filaments bridged the gap between the electrodes. Consequently, the large dynamic range of ON states presented in Figure 3b—namely, from high resistance at low I_C (i.e., 1 MΩ at 100 nA, corresponding to a

switching power < 100 nW), to low resistance at high I_C (i.e., 1 kΩ at 1 mA, corresponding to a switching power of 300 µW)—can be attributed to a modification of the bridging filament morphology, rather than to a modulation of the tunnel barrier length (which is a plausible mechanism in the case of a non-bridging filament).

As a first level of interpretation, the low I_C region can be reasonably described by weak filaments that tend to dissolve very quickly once the voltage is removed. The high I_C region can be considered to correspond to strong bridging filaments with slower relaxation. This effect has been described thermodynamically in Ag filaments [ref27] as a competition between the surface and volume energies: thin filaments tend to be disrupted because the surface energy is higher than the volume energy, whereas thick filaments tend to stabilize because the volume energy is higher than the surface energy. Such relaxation of the conductive paths has been reported in nanoscale devices [ref22,23] and was the basis for the implementation of STP and the STP to LTP transition. After the conductive filament forms via a strong stimulation, the filaments tend to dissolve and the device relaxes toward its insulating state, leading to STP behavior. Stronger stimulation of the device during the SET transition leads to stronger filaments and higher conductance states with more stable characteristics, resulting in LTP. In this case, the conductance state is correlated directly with the volatility.

Assuming that similar dendritic processes occur at the nanometer and millimeter scales (Figure 2a), we can draw a more complex picture for the interpretation of filament stability. Specifically, the different ON states can be described by dendritic trees, in which the resistance is modulated equally by the density and diameter of the branches. At the nanoscale, the same ON state can be obtained by filaments with dense and thin branches as can be obtained by filaments with less dense and thick branches (Figure 3c). Both configurations should lead to different volatilities, emulating different plasticity properties.

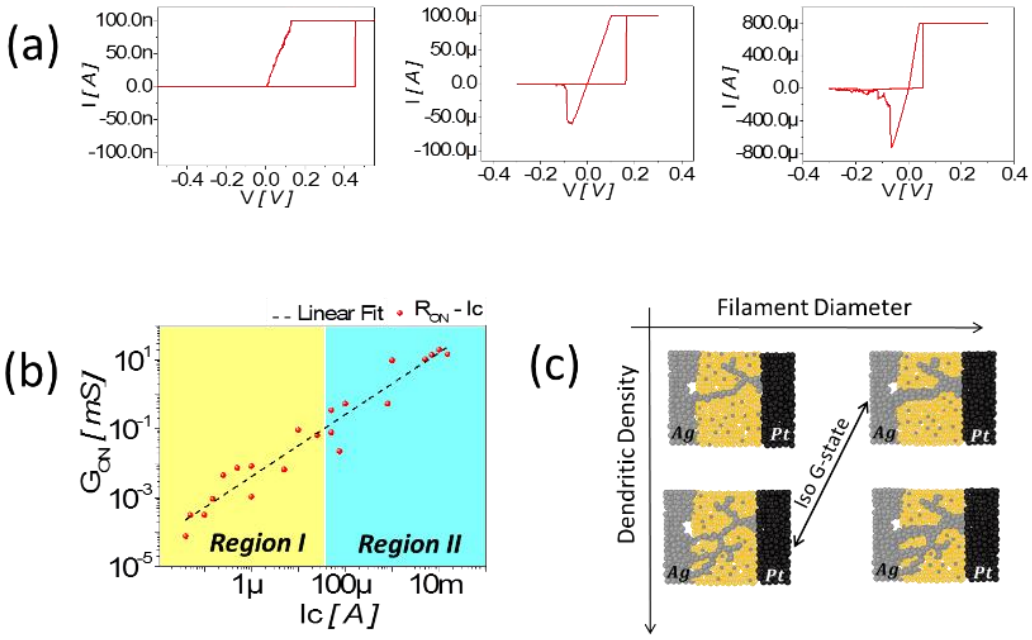

Figure 3: Nanoscale ECM cell configuration. (a) I-V switching characteristics for different values of the compliance current, I_C. When I_C = 100 nA, the ON state is unstable and tends to relax very quickly (OFF transition is not measurable). When I_C = 100 µA or 800 µA, conventional bipolar switching hysteresis loops are obtained, corresponding to the stable ON state. (b) ON state conductance as a function of I_C. Limiting the current during SET limits filament formation. When I_C = 100 nA to 50 µA (region I), the bridging filaments show a high volatility; when I_C > 1 µA (region II), the ON states are stable. (c) Schematic of the proposed scenario describing switching in ECM cells. Both the density and diameter of the dendritic branches can induce an increase in the ON state. The isoconductance state can be obtained with two different filament configurations.

3.2.2. Synaptic plasticity implementation

To evaluate the plasticity properties of our electronic synapses, we performed pulsed measurements with simplified pulses equivalent to the spike rate-coding scheme observed in biological networks. First, a full SET and RESET cycle was realized by voltage sweeping and limiting the current in the SET transition, with the conditioning loop resulting in an initial OFF state equivalent to Figure 3a. Then, the device was exposed to a train of pulses (5 kHz) with fixed amplitude (0.42 V) and width (100 µs), resulting in potentiation of the device (i.e., conductance increase). Relaxation of the synaptic efficiency was sampled over six decades of time by short read pulses with lower voltage (0.1 V) and short duration (100 µs), to minimize the effect on the relaxation mechanism (Figure 4a). Different excitatory bursts, obtained by varying the number of pulses, were used to modulate the potentiation obtained at the end of the pulse sequence, corresponding to the conductance at the end of a burst of pulses, G_{max}. These bursts were fitted by a simple exponential function (Figure 4b).

Consistent with our previous observation that low stability is obtained at a low ON state due to the thinner filaments, we obtained a short relaxation time constant for the lowest ON state. Increasing G_{max} led to a higher time constant and more stable filaments. When we analyzed the evolution of the relaxation time as a function of G_{max} for different I_C values during the conditioning loop (Figure 4c), a second parameter for volatility control emerged. At high I_C values, there was a sharp transition between the low and high time constants. A smoother transition was obtained as G_{max} increased when lower I_C values were used.

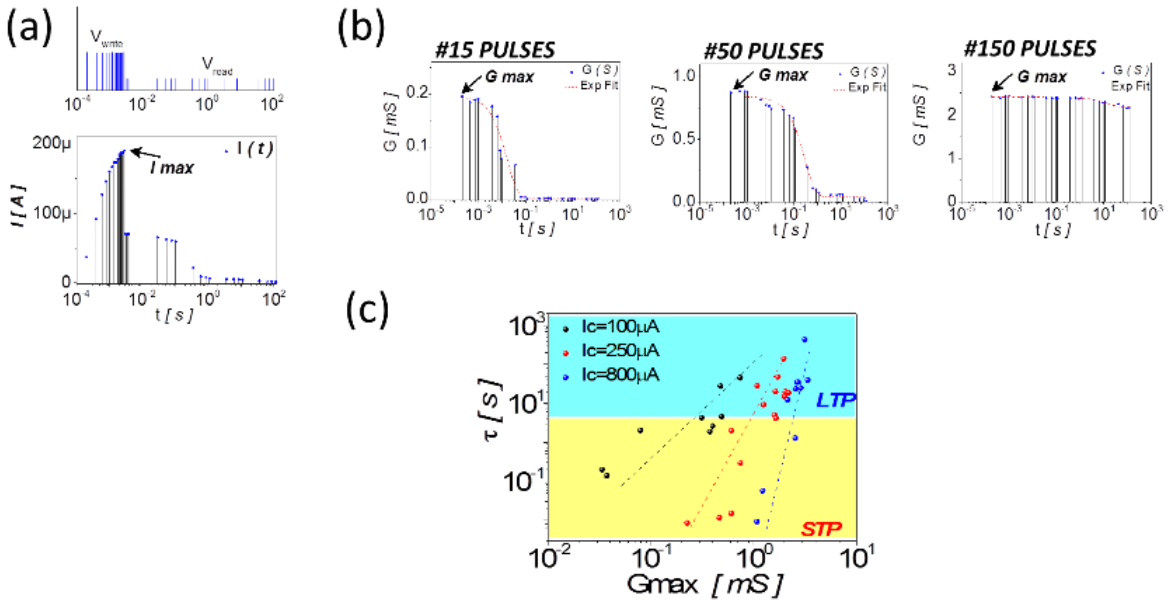

Figure 4: Implementation of the synaptic plasticity. (a) Protocol for the measurement of pulse relaxation. A burst of pulses at 5 kHz and (0.42 V) induced potentiation. Current relaxation was measured at a lower voltage (0.1 V). (b) Measurements of conductance relaxation (blue points) and fitting (red line) on six time decades for different potentiation G_{max} values, obtained by varying the number of pulses (15, 50, and 150 pulses). Low and high G_{max} values led to STP (complete relaxation over time) and LTP (no relaxation over time), respectively. (c) Relaxation time constant as a function of I_C and conductance state at the end of the burst of pulses, G_{max}.

Another formulation of this result is presented in Figure 5a. If we consider the conductance state 100 s after the end of the excitatory burst, then different transitions from STP (relaxation of the conductance state after 100 s; $G_{max} > G_{100s}$) to LTP (no relaxation of the conductance state after 100 s; $G_{max} \approx G_{100s}$, blue area in Figure 5a) can be identified as a function of I_C. This behavior can be attributed to the combination of two effects. Namely, both I_C and the strength of the excitatory burst (i.e., number of pulses) contribute to the definition of the conductive paths. After the conditioning loop, the device is in its OFF state. Traces for the remaining dendritic branches (defined by I_C) correspond to preferential paths for filament formation during the excitatory burst. By analogy with filament formation obtained on millimeter-scale devices, higher I_C

should lead to denser dendritic trees. Thus, the first parameter for plasticity tuning is the I_C value used during conditioning. This value controls the average conductance of the filament during switching in pulse mode, by defining the switching path (i.e., dendrite density). The second parameter that controls the STP to LTP transition is the excitation strength (i.e., number of pulses, which controls G_{max}). This parameter can be associated with an increase of the branch diameter. These two parameters, the past history of the device through the conditioning loops, and the spiking activity during potentiation can be changed independently of each other to modify the device conductance and the filament volatility. Such mechanism is consistent with simulation and experimental results obtained in [ref Pan]. In this work, large filaments are obtained at low surface overpotentials (voltage applied at the electrode/ionic conductor interface) and long switching time while thin filaments results from large surface overpotentials and short switching time. In our case, as the applied voltage is constant, we should consider the voltage redistribution across the full device (i.e. top and bottom interfaces and the ionic conductor's bulk): for low Ic value used during conditioning, the remaining paths correspond to higher bulk's resistivity (low dendritic density) in comparison to high Ic value that leads to denser dendritic paths with lower bulk's resistivity. Consequently, at fixe pulse amplitude, the surface overpotential can be significantly larger in the case of high Ic conditioning loops than in the low Ic regime. In order to reach an equivalent ON state, low Ic conditioning requires a larger switching time (i.e. larger pulse number) leading to large filaments while the high Ic conditioning leads to shorter switching time (i.e. lower pulse number) and thin filaments.

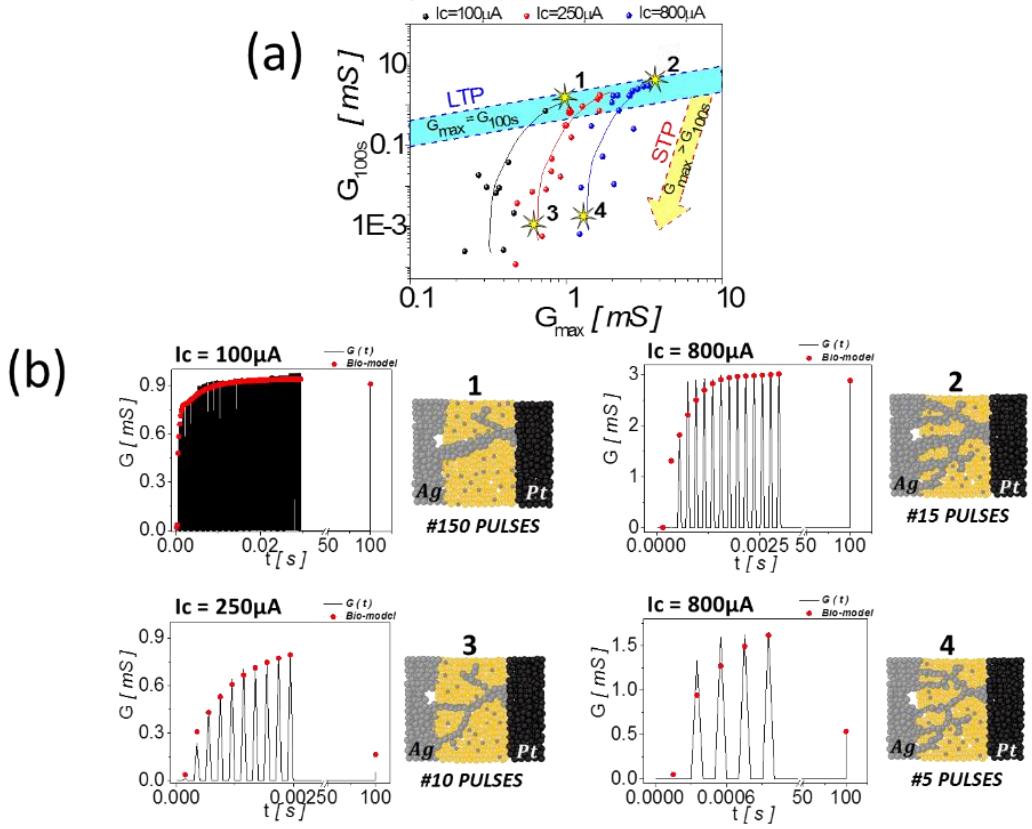

Figure 5: Implementation of the synaptic plasticity. (a) After a conditioning loop (full SET and RESET cycle with current compliance, I_C), the device is stressed with a burst of spikes, which induce a potentiation from the OFF state to a final conductive ON state, G_{max}. Device conductance is measured 100 s after the end of the burst to evaluate the relaxation. Different transitions from STP to LTP are obtained with different conditioning I_C values (I_C = 100, 250, and 800 µA). (b) Two examples of LTP (cases 1 and 2) and STP (cases 3 and 4), for the case in which the number of pulses is set as the key plasticity factor and the I_C value is set as the dendritic path definition. The density (through I_C) and diameter (through burst excitation) of the dendritic branches can be tuned independently to reproduce various STP/LTP combinations.

To illustrate the improved functionality obtained with our approach, we used the biological model of synaptic plasticity developed by Markram [ref28] to fit our different synaptic potentiation experiments (Figure 5b). This model describes the excitatory postsynaptic potentiation response produced by a train of

presynaptic action potentials (APs). After a number of APs (n), the postsynaptic current response to the n+1th AP is given by:

$$I_{n+1} = A_{SE} \cdot R_{n+1} \cdot U_{n+1} \tag{1}$$

where the absolute synaptic efficiency, A_{SE}, corresponds to the maximum possible synaptic efficiency; the fraction of available synapses, R, corresponds to the neurotransmitter resources that are available in the presynaptic connection ($0 < R < 1$); and the utilization of the synaptic efficacy, U, corresponds to the amount of neurotransmitter that is released from the pre- to the postsynaptic connection ($0 < U < 1$). Thus, R_{n+1} and U_{n+1} are given by:

$$\begin{cases} R_{n+1} = R_n(1 - U_{n+1})e^{-\Delta t/\tau_{rec}} + (1 - e^{-\Delta t/\tau_{rec}}) \\ U_{n+1} = U_n e^{-\Delta t/\tau_{fac}} + U_{SE}(1 - U_n)e^{-\Delta t/\tau_{fac}}) \end{cases} \tag{2}$$

The facilitating behavior observed during a burst of spikes is associated with the parameter U_{SE}, which is modified with the characteristic time τ_{fac} and applied to the first AP in a train (i.e., $R_1 = 1 - U_{SE}$). Recovery of the synaptic efficiency (or available neurotransmitters) is associated with the characteristic time τ_{rec}. This biological model allows us to reproduce different kinds of plasticity observed in synapses relative to different mechanisms. Plasticity can be controlled through the neurotransmitter dynamics in the presynaptic connection (i.e., recovery of the available neurotransmitters or increase in the neurotransmitter release probability), by the improvement of neurotransmitter detection in the postsynaptic connection or even by a structural modification of the synaptic connection (i.e., increase in the size of a given synapse or the overall number of synapses connecting two neurons). For a detailed review of synaptic plasticity, see ref. [ref29]. Consequently, the synaptic efficiency of a given spike is determined by a combination of parameters that lead to different synaptic responses and expressions of synaptic plasticity.

Table 1: Fitting parameters used for synaptic plasticity modeling

LTP		**STP**	
case 1: 150 **pulses** $I_c = 100\mu A$	**case 2:** 15 **pulses** $I_c = 800\mu A$	**case 3:** 10 **pulses** $I_c = 250\mu A$	**case 4:** 5 **pulses** $I_c = 800\mu A$
$U_{SE} = 0,0279$	$U_{SE} = 0,0279$	$U_{SE} = 0,0251$	$U_{SE} = 0,0279$
$A_{SE} = 0,0006$	$A_{SE} = 0,0250$	$A_{SE} = 0,0065$	$A_{SE} = 0,0160$
$\tau_{rec} = 0,0013$	$\tau_{rec} = 0,0013$	$\tau_{rec} = 0,0010$	$\tau_{rec} = 0,0012$
$\tau_{fac} = 11,5500$	$\tau_{fac} = 18,5500$	$\tau_{fac} = 0,0150$	$\tau_{fac} = 1,5500$

By accounting for the parameters of the bio-model (Equation 1), four different cases may be analyzed as a function of the number of pulses and I_C (Table 1). If we consider experiments 1 and 3 in Figure 5b, the same potentiation (i.e., G_{max}) can lead to LTP (case 1 with 150 pulses and I_C = 100 µA) or STP (case 3 with 10 pulses and I_C = 250 µA). The STP to LTP transition is mainly associated with an increase of the facilitating time constant, τ_{fac}. This increase is obtained by increasing the number of pulses during the excitatory burst. Slightly increasing I_C is mostly represented by an increase in A_{SE}. This observation is also evident by comparing case 2 with case 4. The difference in conductance level between cases 1 and 2, which showed qualitatively equivalent LTP responses, is mainly attributed to an increase of A_{SE}. We cannot establish a one-to-one correspondence between biological processes (e.g., neurotransmitter dynamics, structural modifications, etc.) and filament growth or relaxation in our experiments because most of the parameters are coupled in

both cases. Additional experiments, such as the in situ observation of filament shape, would provide more insights in order to formulate of more refined equivalence.

3.2. DISCUSSION

Obtaining the synaptic density has been a major challenge in neuromorphic engineering. From a practical perspective, we believe that developing devices that are more functional (i.e., have properties closer to biological synapses) will allow the construction of more complex systems. In a previous report describing the STP to LTP transition [ref14,22], the transition was controlled by a single parameter (i.e., device conductance). Such behavior was proposed as a direct solution for the implementation of the multistore memory model [ref] which considers that learning events contribute to the formation of short term memory (where memory is used in the sense of psychology) before being transferred into long term memory (STM/LTM transition). If a direct equivalence between STP/LTP and STM/LTM is not straightforward, it seems realistic to consider synaptic plasticity as a key element in the formation of memory. The device presented in this paper features a tunable STP/LTP transition that could be a key parameter for defining the appropriate activity threshold that determines when information storage needs to be moved from a short term to a long term regime, or, in other words, how long an information needs to be sustained (i.e. how long the device will remain in its ON state).

Additionally, if STP/LTP transition is only controlled by the device's conductance, synaptic weight modification and STP/LTP transition cannot be uncorrelated. We argue that the rate-coding property obtained in the STP regime, as observed in the facilitation of synaptic signal transmission during a high frequency burst of spikes and the subsequent relaxation at lower frequencies, disappears once the device enters into its LTP regime and, thus, becomes a linear resistor. From a circuit perspective, if we consider a simple integrate-and-fire neuron associated with linear synapses, the node (neuron and synapses) is equivalent to a simple linear filter (if the variable is the average spiking rate). The node is a nonlinear filter in the STP regime with frequency-dependent synaptic conductance. The overall network functionality is reduced when learning moves synapses from their STP to their LTP domain. An interesting property offered by the presented devices in order to preserved such rate coding functionality is to allow for weight modification through the control of the A_{SE} parameter while maintaining the frequency dependent response by keeping the device into its short term regime (see case 3 and 4, figure 5). For the device presented in this paper, learning can be realized by modifying the dendritic filament density and increasing the A_{SE} during the conditioning procedure. The frequency coding property can be ensured by controlling the filament diameter and relaxation.

Finally, the activity dependent STP/LTP transition and synaptic weight modification in this work is only obtained as a function of the input frequency, thus corresponding to the pre-neuron activity. Such mechanism is defined in biology as a facilitating synapse. A complementary mechanism that cannot be reproduce with our system is the depressing synapse (i.e. decrease of the synaptic weight when pre-neuron activity increase). In order to implement practical learning systems, this results will have to be extended to hebbian learning strategies in which weight modification is dependent on both pre- and post-neuron activity. Among the different hebbian learning strategies considered to date, STDP has attracted a large attention. One implementation of such learning protocol is based on overlapping pulses (spike timing difference between pre- and post-neuron is then encoded as a voltage drop across the device). Figure SXXX presents similar results to figure 5 when voltage is used as a key plasticity factor instead of spiking frequency that should allow for STDP realization. While not measured in this paper, one interesting future direction would be to add to previously report STDP results obtained on non-volatile systems [ref Jo] the STP/LTP capacity in order to demonstrate neuromorphic circuits with richer dynamical behaviors.

3.3. CONCLUSIONS

We report a single synaptic device that highly resembles its biological counterpart, opening the field to more complex neuromorphic systems. Biological synaptic plasticity has been successfully implemented in our nanoscale memristive device by considering the filament stability of ECM cells, in terms of competition between the density and diameter of the dendritic branches. STP and LTP regimes can be controlled by tuning the device volatility. The first parameter for plasticity tuning, I_C, is used during conditioning and controls the average conductance of the filament during switching in pulse mode. The second parameter handles the STP to LTP transition and corresponds to the excitation strength (number of pulses), which controls G_{max}. The second parameter can be associated with an increase of the branch diameter. These two parameters can be tuned independently of each other to modify the device conductance and filament volatility. Future work should investigate how such synaptic properties can be advantageous for large-scale neuromorphic circuits. To improve the efficiency of future bio-inspired computing systems, interdisciplinary research is needed to obtain a better understanding of the contributions of STP and LTP mechanisms to memory construction and spike-coding information processing.

1. Merolla, P. A.; Arthur, J. V.; Alvarez-Icaza, R.; Cassidy, A. S.; Sawada, J.; Akopyan, F.; Jackson, B. L.; Imam, N.; Guo, C.; Nakamura, Y. A Million Spiking-Neuron Integrated Circuit with a Scalable Communication Network and Interface. Science 2014, 345, 668–673.
2. Alibart, F.; Zamanidoost, E.; Strukov, D. B. Pattern Classification by Memristive Crossbar Circuits Using Ex Situ and In Situ Training. Nat. Commun. 2013, 4, 403–405.
3. Strukov, D. B.; Snider, G. S.; Stewart, D. R.; Williams, R. S. The Missing Memristor Found. Nature 2008, 453, 80–83.
4. Liang, B.; Dubey, P. Recognition, Mining and Synthesis. Intel Technol. J. 2005, 9, 99–174.
5. Pickett, M. D.; Medeiros-Ribeiro, G.; Williams, R. S. A Scalable Neuristor Built with Mott Memristors. Nat. Mater. 2013, 12, 114–117.
6. Sharad, M.; Augustine, C.; Panagopoulos, G.; Roy, K. Spinbased Neuron Model with Domain-wall Magnets as Synapse. IEEE Trans. Nanotechnol. 2012, 11, 843–853.
7. Strukov, D. B. Nanotechnology: Smart Connections. Nature 2011, 476, 403–405.
8. Markram, H.; Lübke, J.; Frotscher, M.; Sakmann, B. Regulation of Synaptic Efficacy by Coincidence of Postsynaptic APs and EPSPs. Science 1997, 275, 213–215.
9. Hebb, D. The Organization of Behavior: A Neuropsychological Theory; Psychology Press: New York, 1949; pp 4356.
10. Bi, G.-q.; Poo, M.-m. Synaptic Modifications in Cultured Hippocampal Neurons: Dependence on Spike Timing, Synaptic Strength, and Postsynaptic Cell Type. J. Neurosci. 1998, 18, 10464–10472.
11. Jo, S. H.; Chang, T.; Ebong, I.; Bhadviya, B. B.; Mazumder, P.; Lu, W. Nanoscale Memristor Device as Synapse in Neuromorphic Systems. Nano Lett. 2010, 10, 1297–1301.
12. Kuzum, D.; Jeyasingh, R. G.; Lee, B.; Wong, H.-S. P. Nanoelectronic Programmable Synapses Based on Phase Change Materials for Brain-inspired Computing. Nano Lett. 2011, 12, 2179–2186.
13. Choi, S.-J.; Kim, G.-B.; Lee, K.; Kim, K.-H.; Yang, W.-Y.; Cho, S.; Bae, H.-J.; Seo, D.-S.; Kim, S.-I.; Lee, K.-J. Synaptic Behaviors of a Single MetalOxideMetal Resistive Device. Appl. Phys. A 2011, 102, 1019–1025.
14. Kim, K.; Chen, C.-L.; Truong, Q.; Shen, A. M.; Chen, Y. A Carbon Nanotube Synapse with Dynamic Logic and Learning. Adv. Mater. 2013, 25, 1693–1698.
15. Zeng, F.; Li, S.; Yang, J.; Pan, F.; Guo, D. Learning Processe Modulated by the Interface Effects in a Ti/conducting polymer/Ti Resistive Switching Cell. R. Soc. Chem. Adv. 2014, 4, 14822–14828.
16. Krzysteczko, P.; Münchenberger, J.; Schäfers, M.; Reiss, G.;Thomas, A. The Memristive Magnetic Tunnel Junction as a Nanoscopic Synapse-Neuron System. Adv. Mater. 2012, 24, 762–766.
17. Alibart, F.; Pleutin, S.; Bichler, O.; Gamrat, C.; Serrano-Gotarredona, T.; Linares-Barranco, B.; Vuillaume, D. A Memristive Nanoparticle/Organic Hybrid Synapstor for Neuroinspired Computing. Adv. Funct. Mater. 2012, 22, 609–616.
18. Alibart, F.; Pleutin, S.; Guérin, D.; Novembre, C.; Lenfant, S.; Lmimouni, K.; Gamrat, C.; Vuillaume, D. An Organic Nanoparticle Transistor Behaving as a Biological Spiking Synapse. Adv. Funct. Mater. 2010, 20, 330–337.
19. Zhu, L. Q.; Wan, C. J.; Guo, L. Q.; Shi, Y.; Wan, Q. Artificial Synapse Network on Inorganic Proton Conductor for Neuromorphic Systems. Nat. Commun. 2014, 5, 1–7.

20. Josberger, E. E.; Deng, Y.; Sun, W.; Kautz, R.; Rolandi, M. Two-Terminal Protonic Devices with Synaptic-Like Short-Term Depression and Device Memory. Adv. Mater. 2014, 4986–4990.
21. Kim, S.; Choi, S.; Lu, W. Comprehensive Physical Model of Dynamic Resistive Switching in an Oxide Memristor. ACS Nano 2014, 8, 2369–2376.
22. Ohno, T.; Hasegawa, T.; Tsuruoka, T.; Terabe, K.; Gimzewski J. K.; Aono, M. Short-Term Plasticity and Long-Term Potentiation Mimicked in Single Inorganic Synapses. Nat. Mater. 2011, 10, 591–595.
23. Yang, R.; Terabe, K.; Liu, G.; Tsuruoka, T.; Hasegawa, T.; Gimzewski, J. K.; Aono, M. On-Demand Nanodevice With Electrical and Neuromorphic Multifunction Realized by Local Ion Migration. ACS Nano 2012, 6, 9515–9521.
24. McGaugh, J. L. Memory;a Century of Consolidation. Science 2000, 287, 248–251.
25. Valov, I.; Waser, R.; Jameson, J. R.; Kozicki, M. N. Electrochemical Metallization Memories Fundamentals, Applications, Prospects. Nanotechnology 2011, 22, 254003.
26. Valov, I.; Linn, E.; Tappertzhofen, S.; Schmelzer, S.; van den Hurk, J.; Lentz, F.; Waser, R. Nanobatteries in Redox-Based Resistive Switches Require Extension of Memristor Theory. Nat. Commun. 2013, 4, 1771.
27. Russo, U.; Kamalanathan, D.; Ielmini, D.; Lacaita, A. L.; Kozicki, M. N. Study of Multilevel Programming in Programmable Metallization Cell (PMC) Memory. IEEE Trans. Electron Devices 2009, 56, 1040–1047.
28. Hsiung, C.-P.; Liao, H.-W.; Gan, J.-Y.; Wu, T.-B.; Hwang, J.-C.; Chen, F.; Tsai, M.-J. Formation and Instability of Silver Nanofilament in Ag-Based Programmable Metallization cells. ACS Nano 2010, 4, 5414–5420.
29. Pan, F.; Yin, S.; Subramanian, V. A Detailed Study of the Forming Stage of an Electrochemical Resistive Switching Memory by KMC Simulation. IEEE Electron Devices Lett. 2011, 32, 949–951.
30. Markram, H.; Pikus, D.; Gupta, A.; Tsodyks, M. Potential for Multiple Mechanisms, Phenomena and Algorithms for Synaptic Plasticity at Single Synapses. Neuropharmacology 1998, 37, 489–500.
31. Zucker, R. S.; Regehr, W. G. Short-Term Synaptic Plasticity. Annu. Rev. Physiol. 2002, 64, 355–405.
32. Collingridge, G. L.; Peineau, S.; Howland, J. G.; Wang, Y. T. Long-Term Depression in the CNS. Nat. Rev. Neurosci. 2010, 11, 459–473.
33. Atkinson, R. C.; Shiffrin, R. M. Human Memory: A Proposed System and its Control Processes. Psychol. Learn. Motiv. 1968, 2, 89–195.

4. CHAPTER 4

Neuromorphic Time-Dependent Pattern Classification with Organic Electrochemical Transistor Arrays

4.1. INTRODUCTION

In biological systems, dynamical and complex information is processed efficiently by highly redundant and parallel network of cells while standard computing systems are quickly reaching their limitations for equivalent information processing tasks. For instance, at the opposite to top-down circuits with highly uniform devices used for general purpose computers, bottom-up assembly of neural cells with high level of variability, can process auditory, visual or olfactive stimuli and generate complex actions very efficiently. Material implementation of such bio-inspired principles for sensing and computing has been a stimulating direction that has reached significant milestones with the development of neuromorphic sensors (retina, cochlea…) and circuits.[1] While initially relying on standard silicon-based devices (i.e. CMOS), emerging materials and devices are opening new avenues for neuromorphic engineering by offering new basic mechanisms for emulating biology and new devices and circuits concepts to build computing systems. Notably, neuromorphic systems with non-volatile memories (and resistive memory in particular) have been the focus of strong research efforts.

Here, we capitalize on organic electro chemical transistors (OECTs) that have been recently proposed ubiquitously as basic building blocks in neuromorphic computing applications (i.e. memory devices, for instance) and as bio-sensors thanks to their intrinsic sensitivity to ions. While this paper does not present sensor properties assessment, we build on this intrinsic feature of OECT to demonstrate neuromorphic computing application. Based on an array of organic ionic transistors, this work shows how sensing and processing can be realized at the interface with an analyte by taking advantage of OECTs intrinsic physics and on neuromorphic concepts. In particular, we show how highly variable material engineering routes that are not adapted to standard information processing technologies (i.e. relying on top-down fabrication of near-ideal components and circuits) can be turned into an advantage when bio-inspired concepts are used to engineer computing system.

We propose an adaptation of the recent proposition of reservoir computing (RC),[2,3] to demonstrate that both sensing and computing can be obtained from the intrinsic properties of a transistor array, limiting the separation between these two elementary levels (i.e. sensing and computing). In one hand, from the neuromorphic computing side, RC concept has been developed for dynamical signal processing (e.g. speech recognition),[4] and use the idea of learning from a simple read-out layer (i.e. a feedforward perceptron) the dynamics associated to the projection of a given stimuli into a complex and random network of non-linear elements (i.e. neurons or nodes). In the other hand, from the sensing perspective, monitoring and analyzing biological activity in medium such as neural cells assembly or bloods composition, for instance, consist in processing dynamical signals and would strongly benefit from the RC approach to classify such dynamical patterns from complex and poorly define biological medium. Thus developing RC strategies for processing information out of a network of ion-sensitive transistors could open new perspectives in biological sensors. The key elements of RC are (i) non-linearity of the reservoir's nodes (non-linear conversion from input signal to output signal) and (ii) a fading memory effect keeping the history of the stimuli active in the network on a given duration. Material implementations of RC have been proposed recently with optical or magnetic oscillators.[5-9] For both, time multiplexing was used in order to emulate spatial nodes in the network from

one single non-linear element and memory effect was associated whether to a feedback loop connection or to the transient dynamics of the non-linear element. Here, we propose the implementation of a spatial reservoir composed of an array of OECTs that present a non-linear response to the stimulus propagating in an analyte (an input voltage applied to the analyte is converted into a resistance state of the OECT. The memory effect is associated to the transient dynamics of ions penetrating into the OECT with a given relaxation time implementing the fading memory. We show in this paper that this spatial reservoir take advantage of (i) the variability in the OECTs array inherent to the bottom up fabrication of the OECTs using a newly synthesized electropolymerizable polymer, (ii) of the transient dynamics of the devices for an implicit representation of time and (iii) of the number of redundant OECTs to discriminate simple dynamical patterns.

4.2. RESULTS

4.2.1. Transient dynamics of OECTs as implicit time representation

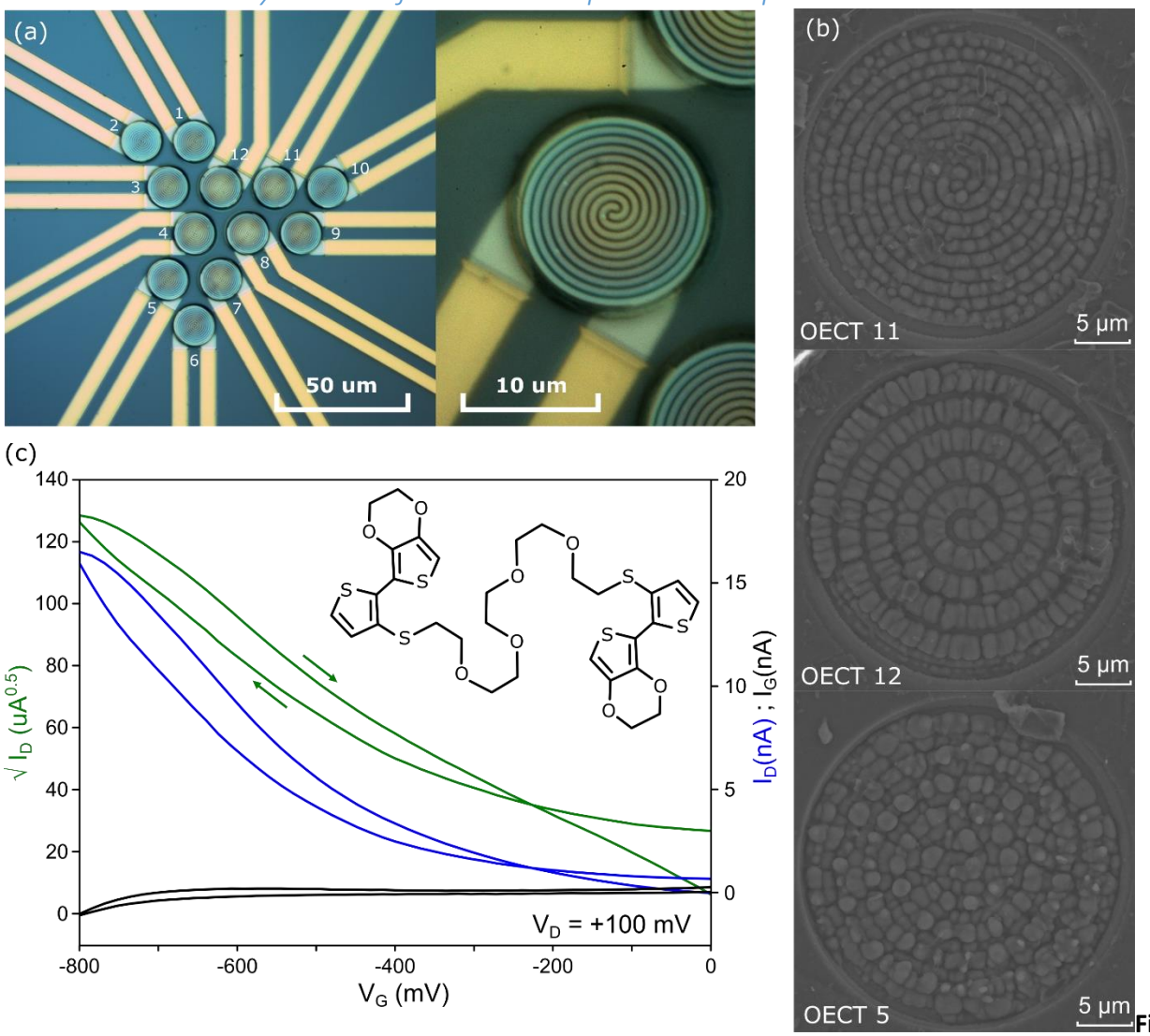

Figure 1. a, Optical micrographs of the OECTs array. OECT devices are three terminal devices with spiral-shape S and D electrodes leading to a large channel width over length ratio (W/L=1100) over a confined area (615 µm²). The gate electrode (not represented) is realized with macroscopic metal wire contacting the electrolyte. b, After electrode patterning, the organic material, Scanning Electron Microscope images of three different devices. High variability in the electro-polymerization is apparent in the polymer structure. c, Transfer characteristics of one of the OECT in a 0.1 M $KCl_{(aq)}$ electrolyte, displaying the p-type accumulation-mode field-effect (inset: TEDOT monomer used for the semiconductor electrodeposition). Blue line: drain current, green line: square-root of drain current, black line: gate current.

The OECT array (**Figure 1**a, and Methods) consists in organic electrochemical transistors composed of electro-polymerized, glycol crosslinked, 2-(2-thienyl)-(3,4-ethylenedioxythiophene) (TEDOT) molecules (Figure 1 and

Supplementary Figure S1). Alternative materials based on glycol-side-chain polythiophene have demonstrated high performances while operating in accumulation mode.[13,14] We used the monomer TEDOT (Figure 1) to conceive, after electro-polymerization, a new polythiophene functionalized with ethylene glycol chains, patterned locally on each OECT. After the polymer electrodeposition (see Methods), all devices showed gate modulation of source-drain current (Figure 1c), despite the large variability of the material morphology (Figure 1b). The polymer thickness could not be estimated by profilometry due to this textured morphology of all devices. From the consideration of a materials density of 2 g/cm^{-3}, the dimensions of the device cavities and the charge flow controlled over the potentiostatic deposition, we evaluated a thickness of a theoretical thin film of 6 µm with seems to agree with the experiment (practically the polymer patches fill the Parylene C cavities). The basic mechanisms of OECT is based on the redox doping/dedoping of the organic material.[15,16] A negative gate voltage (V_G) applies to the 0.1 M $KCl_{(aq)}$ (the analyte) forces negative ions to penetrate into the organic material, increasing the electronic conductance of the organic layer (source grounded and drain potential constant V_D=+100 mV). When V_G is turned off, ions diffuse back to the electrolyte, out of the organic material that recovers its high resistance state. Non-linear relationship between V_G and device's resistance is evident from Figure 1b. In addition, slow dynamics of ions through the electrolyte/organic interface is apparent in the hysteresis loop when sweeping V_G at 0.1 V/s. **Figure 2**a presents the transient response of an OECT to a sequence of pulses with increasing amplitude. The transient behavior of the OECT is used to implement short-term memory effect (Figure 2c),[17,18] as observed in biological synapses (Short-Term Plasticty).[19] When the OECT is stressed with a train of pulses of constant amplitude, short-term facilitation (increase of the average output current with number of pulse) is implemented.[17,20]

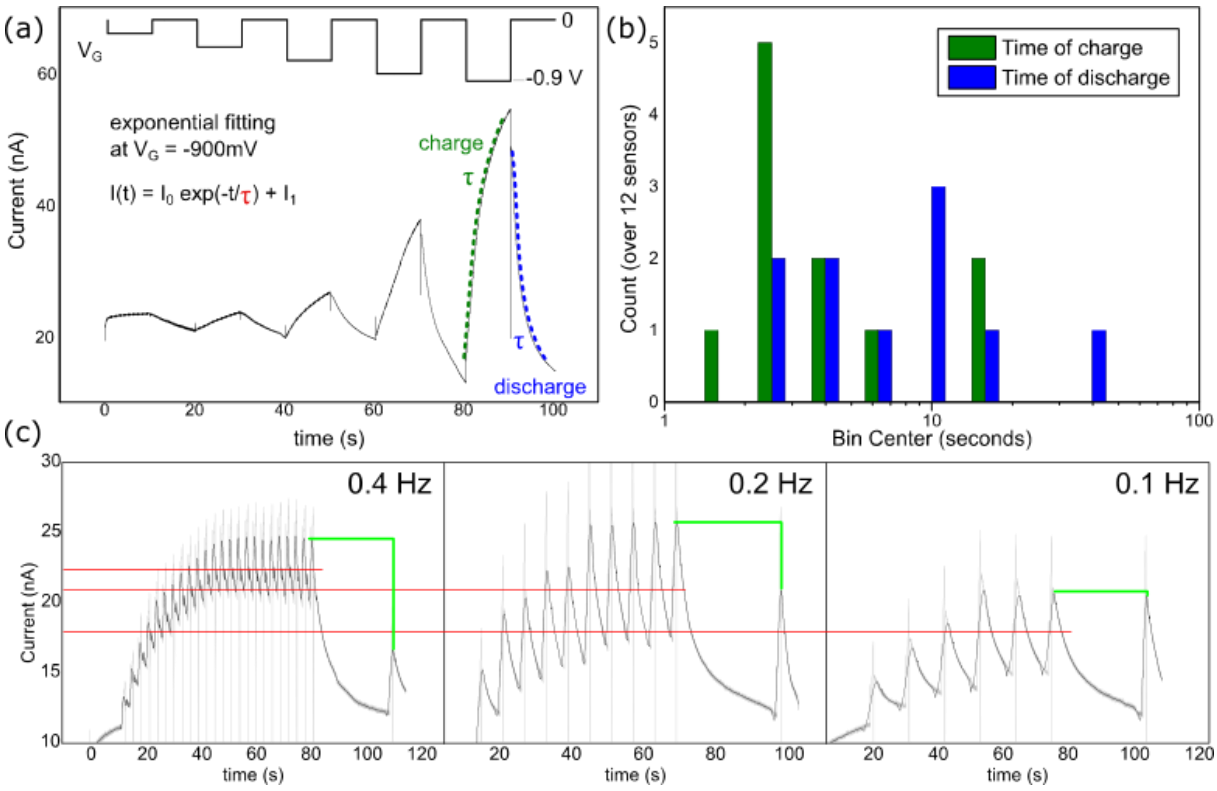

Figure 2. a, Typical response of the OECT to a gate voltage with different amplitudes and constant V_D=+100mV. Green (blue) dashed line shows the fitting of the doping (dedoping), transient resulting from ions injected (removed) to (from) the organic semiconductor for a step voltage of 0 to -900 mV (-900 mV to 0 V). b, Characteristic charge and discharge time constants for the full array of 12 OECTs obtained from fitting the transient current in (a) by exponential functions (dashed lines in (a)). c, Constant gate voltage stimulation (-900 mV, 200 ms width) with variable frequencies showing short-term facilitation. Red dashed lines correspond to the average conductance reach in steady-state showing frequency-dependent potentiation. Short-term memory effect is evidenced by the single pulse applied 25 seconds after the potentiation. Green lines are guide to the eyes to evidence the stronger potentiation/relaxation ratio after higher frequency stimulation.

This short-term memory effect is due to the accumulation of ions from pulse to pulse and on the unbalanced relaxation between each stimulation resulting in higher steady-state mean conductance at high frequency (larger amount of ions are accumulated) and lower steady-state conductance at low frequency (lower amount of ions are accumulated). In each case in Figure 2c, a single control pulse is applied after 25 s of relaxation to evidence the short-term memory effect. The absolute value of the current from trial-to-trial presents some variability in current level (evidenced in the very first pulse when the OECT was previously at rest, for instance) that was mostly due to a poor control of the electrolyte concentration from trial-to-trial and to the strong effect of the concentration on the quantitative response of the OECT. Nevertheless, the qualitative response showing higher potentiation at higher frequency with respect to the control pulse (green lines in Figure 2c) remained consistent. The transient behavior was analyzed by fitting the charge/discharge in the transient drain current characteristics of Figure 2a with a single exponential function. Characteristic times with a large dispersion are obtained (Figure 2b), inherent to the bottom-up fabrication process of the organic material. Characteristic transient times in an OECT depend on the electrical resistance of the polymer and the resistance and capacitance values of the electrolyte [Bernards], affecting the device behavior under pulse modulation [Dual Sensing]. Since the polymer and electrolyte resistances as well as the device-to-electrolyte capacitance are intrinsically function of the thickness of the polymer materials and the areas of their interfaces, the variability of polymer morphology (as shown in Figure 1b) is the source of variability of the measured device time constants. This will be one central element that we exploit in the following for the implementation of RC. Characteristic time constants for charging are on average shorter (from 1.08 to 13.7 s) than discharging (from 2.03 to 43.7 s). The level of memory of each individual device can be define along two metrics: (i) τ_{mean} , the average value of $\tau_{discharge}$ and τ_{charge} and (ii) the $\tau_{discharge}$ / τ_{charge} ratio. When $\tau_{discharge}$ / $\tau_{charge} \approx 1$ (equivalent to a capacitor), the device is a purely short-term memory. When $\tau_{discharge}$ / τ_{charge} tends to higher values, the memory moves from short-term to long-term memory (note that non-volatile memory tends to maximize this ratio with $\tau_{discharge} > 10$ years and $\tau_{charge} < 1$ ns). Figure S5 and S6 represents the $\tau_{discharge}$ / τ_{charge} ratio and τ_{mean} showing that OECTs devices are in the short term regime of memory. In the following, this short term memory effect will be used to define the global memory of our system. The collection of τ_{mean} available in our system due to variability will be used to reconstruct a memory time window. This characteristic memory time window directly determines the typical duration of dynamical patterns that can be processed by the reservoir of OECTs (i.e. a device can keep memory of its previous history on a time window of up to tenths of seconds). In its initial version, RC used recurrent connections into the reservoir to implement fading memory effect. Feedbacks (i.e. delays) into the reservoir ensure that signals are kept for a given time active in the network. Also, strength of this recurrent connections was used to set the reservoir in an optimal state in terms of sensitivity to input signal (i.e. edge of chaotic regime).[21] In our case, the reservoir consists in a purely feed-forward network (i.e. no recurrent connections) and fading memory effect is implemented with the transient responses of OECTs (more precisely by the collection of time constant from each individual OECT). The optimal V_G range of operation of the OECTs is defined based on the device $I_D(V_G)$ characteristics. Too large voltage biases (>1.0 V) might lead to irreversible material damage by water electrolysis,[22] hindering the stability of the electrochemical system. Too small voltages result in too weak modulation of the conductance. As a trade of, we use voltage pulses V_G of -900 mV.

4.2.2. Reservoir computing: dynamical signal processing with network of OECTs

Various Artificial Neural Networks (ANNs) approaches have been proposed so far for time-dependent pattern classification (speech, for example). Recurrent networks or Time Delay Neural Network are of particular interest for this task since they offer the possibility to encode the time signature of such signals explicitly. More recently, time-dependent signals processing has been revitalized with the concept of Reservoir Computing (RC). RC is based on the basic idea of projecting the input signal on the nodes of a large dimensional space in order to separate simple features from the input signal. These simple features are then used to classify patterns at the read-out layer (i.e. a simple perceptron trained with standard learning technics). In time-multiplexed RC approaches used for speech recognition,[5,8] the dimensionality of the reservoir is ensured by the virtual neurons (i.e. virtual nodes) that hold the signature of the signal at different time intervals. Reconstruction on the read-out layer of the time-dependent signals out of these virtual nodes is then used to classify patterns (i.e. speech signals). This approach corresponds to an explicit representation

of time where the first neuron is associated to the first time interval, second neuron to the second time interval… and so on.[23] Here, we use an implicit representation of time through the transient dynamics of each OECT in the network (**Figure 3**a). Due to the variability in their transient responses (Figure 2), each OECT will keep the temporal signature of the signal on a different memory window. Each OECT is then used to collect different features from the reservoir and to perform classification at the read-out layer.

To test this concept, we designed low complexity signals consisting in square waves of constant amplitude -900 mV and 1 s duration applied to the global gate with variable frequencies. The two signals used to demonstrate time-dependent signal classification are built with square-type and triangle-type pulse-frequency modulation between 0.3 and 0.8 Hz.[24] If classification of these two signals is trivial when one have access to the full recording over a complete period of the signals, discrimination of the two signals on a restricted time interval (typically no more than two successive pulses) becomes impossible without some memorization of the past events. Here we show that the RC concept can be used to classify in real time these signals based on the intrinsic memory of each OECT and on pre-requisite learning. Figures 3b-c present the typical response of an OECT to the two signals, respectively. Light grey lines correspond to the as recorded signal. The blue and green lines correspond to the average current response in each time step.

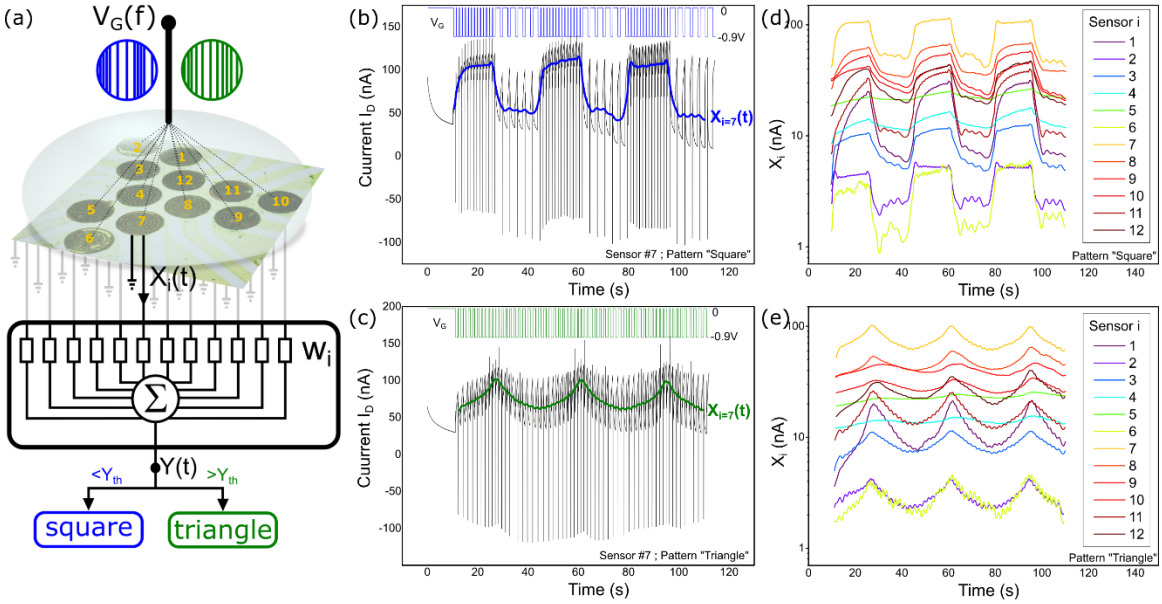

Figure 3. a, Schematic of the experiment and analysis. The network of 12 OECTs is used to sense the response of the electrolyte to the global gate signal. The perceptron is implemented in software and realize signal weighting, summation and activation function. b and c, Typical response of an OECT to the two different patterns (square and triangle, respectively) used for classification. Black/grey lines are the raw data measurement and blue/green line are smoothed signals corresponding to the average current value sense by the device. As in Fig. 2c, highest (lowest) frequencies tends to accumulate more (less) ions in the organic material and increase (decrease) the mean conductance levels of the OECT via doping effect. d and e, Response of the 12 OECTs to the two input patterns (square type and triangle type, respectively). We observe variability from device to device on the modulation amplitude and the shape of the mean current. Average currents of OECT #6 and #7 differ of about two orders of magnitude, and amplitude of the current modulations between OECT #1 and #4 differ of about one order of magnitude.

Figures 3d-e present the responses of the 12 OECTs in the array to the same signal applied at the common gate. We observe variability from device to device on the modulation amplitude and the shape of the mean current. The observed variability on the mean current is not only due to the device time constant variability, but also to the device steady state current and its current modulation. Both correlated to the hole conductivity of the polymer between source and drain electrodes, these two features being highly influenced by the polymer morphology (Figure 1b shows very characteristic polymer grain boundaries between three OECTs with different base currents and current modulations displayed in Figure 3d-e). This large variability represents a severe limitation of bottom-up fabrication technics that RC can leverage efficiently. Implementation of the reservoir target classification of the two patterns (square-type and triangle-type) with

a simple read-out equivalent to a simple perceptron implemented here in software (i.e. one neuron with m=12 weighted input). Output current from each OECT is sampled over time and will be use to define the state of the reservoir at each time step, for a given location into the reservoir. The collection of outputs from the array of OECTs correspond to the state at time t of the reservoir in response to a given stimuli. This output values are then used as input to a simple perceptron in charge of classification through learning (i.e. the perceptron function is limited to signal weighting, summation and activation function). We use a 1 ms sampling rate of the signals. For each time step, we associate a given vector $\{X_i\}(t)$ of dimension (1x12). A total of 11750 vectors are recorded from each pattern composed of three repetitions of one period of square (triangle, respectively) elementary pattern. Each vector is then fed to a simple perceptron with m=12 weighted inputs. The total output Y(t) from the perceptron before activation function at time t is then

$$Y(t) = \sum_{i=1}^{m=12} w_i \cdot X_i(t) \qquad (1)$$

with w_i the synaptic weight of the i^{th} input line, $X_i(t)$ the current value at time t of OECT #i. Vectors $\{X_i\}$ belongings to the triangle-type pattern are associated to class "1" (i.e. output neuron is activated) and square-type pattern to class "0" (i.e. output neuron is not activated). The activation function rule of the output neuron is then

Output = "1" if $Y(t) > Y_{th}$

Output = "0" if $Y(t) < Y_{th}$

with Y_{th} the threshold activation value (O.5 in our example) chosen based on the distribution of the different output values. We define the training vectors (or training examples) by choosing randomly n vectors from each pattern. Training protocol is realized with pseudo-inverse learning (i.e. Moore Penrose operator)[25] to determine the value of the 12 synaptic weights (Supplementary Figure S2). Testing is realized on the full set of vectors (i.e. 11750) from the two classes. Classification performances are then calculated as the percentage of errors averaged on 20 iterations with 2n training vectors randomly chosen among the entire set. **Figures 4**a-b show the error rate as a function of training vectors used for learning the synaptic weights. As in standard learning technics, more examples lead to better performances. With more than n=300 vectors (only 2.6% of the total vectors set), the system reaches classification of the signal with an error rate at the order of 0.001%.

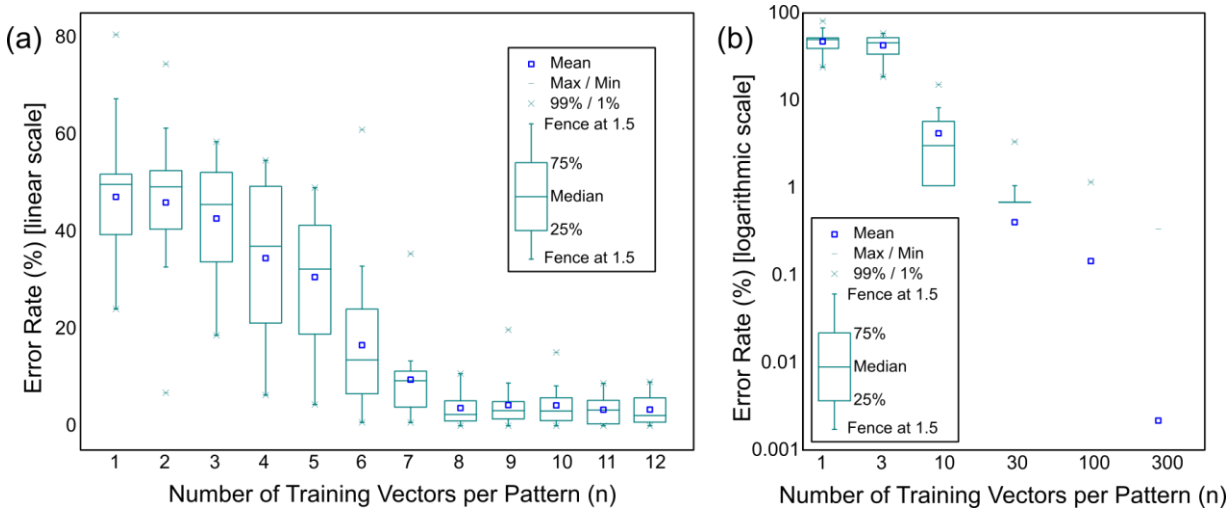

Figure 4. Error rate (i.e. the number of vectors given a mistaken recognition over the total number of tested vectors) for classification of the two patterns as a function of number of training vectors n in each class. Testing is performed on the entire vector set (11750 vectors). a, linear scale. b, logarithmic scale.

The present classification task is a toy problem and any quantitative discussion on the error rate is pointless. The important aspect here is to show that the system after learning can classify in real time input signals since the only operation to be realized out of the OECT array is current weighting, summation and thresholding (excluding the learning stage which require a significant initial computing power and time). Note

that we used the average value of current that can be easily implemented in hardware with a simple integration circuitry.

These results are based on two important aspects. (i) The number of OECTs used for pattern classification that we have associated to the number of features collected from the reservoir. (ii) The variability from device to device that affects how much each feature is different from the other. For instance, two identical OECTs will provide the same feature while very different OECTs will provide very different features (i.e. different transient time between two devices will provide different memory window and consequently, different type of features). To demonstrate that RC takes directly advantage of both number of features and variability in the OECT array, we realize the same classification task with only "partial" arrays of m OECTs, with $1 \leq m \leq 12$.

4.2.3. Influence of the number of OECT in the reservoir

Performance of classification should be directly linked to the number of features (associated to each OECT) used to classify the patterns. To test this hypothesis, we evaluate the performances of the reservoir, degrading it on-purpose by removing sequentially OECTs one by one. **Figures 5**a-b present the error rate as a function of the number of training vectors n when one OECT is removed from one batch to the other.

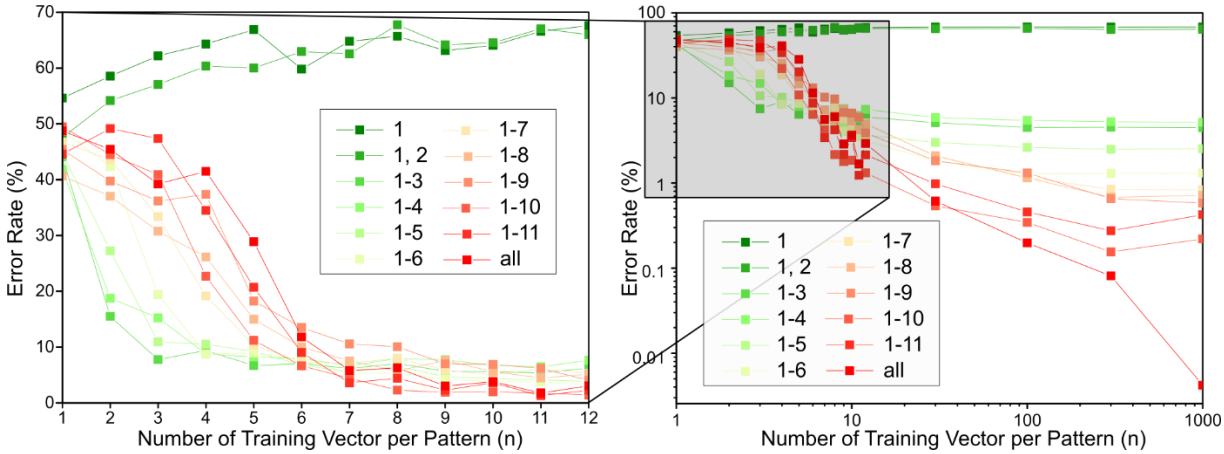

Figure 5. Error rate of the classification task when the OECTs' array is progressively degraded by removing sequentially OECTs one-by-one.

A minimum of three OECTs is required to classify patterns. When more than 300 training vectors are used, a clear relationship between number of OECTs and performances is extracted with the later improving exponentially with the number of OECT used to classify the pattern. Performances of classification are directly linked to the number of OECT used to compute the pattern out of the reservoir, or equivalently, to the number of features used to describe the patterns. Referring to the general idea of RC, increasing the number of OECT in the network corresponds to increasing the dimensionality of the space used to project the input signal. The expected effect is consequently to ease the readout layer (i.e. the perceptron) to classify the patterns. In this sense, more OECTs are better for classifying time-dependent signals. Classification of more complex patterns (i.e. patterns with higher level of similarity, for example) should require more OECTs in the array. We notice that small arrays ($m<6$) require less training vectors to reach error rate of about 10%. In agreement with theoretical prediction from ref. [26], this effect implies that for a relatively small number of training vectors, it exists an optimal number of features to reach the best accuracy. In other words, it shows that the training shall be adapted to the size of the OECT array, which should be adapted to the task complexity.

4.2.4. Influence of the variability in the reservoir

Another important issue is to know what type of features could lead to better performances. **Figure 6**a shows the error rate for the particular case of 6 OECTs out of 12 when all combinations (i.e. 924) are tested. Figure 6a clearly shows that some set of OECTs perform better than other, but it was not possible to extract a possible empirical rule correlating the error rate to the level of variability in transient dynamics and/or mean current level (see Supplementary Figure S4 for details). Nevertheless, important insights are obtained by considering the average performances of the array as a function of number, m, of OECTs. Figures 6b-c show

the error rate for different sizes of array (m) obtained with n=3 (small training set) and n=300 (large training set) training vectors. Each value is obtained by extracting the mean error rate from 10 randomly chosen sets of OECTs (the same sets for both n=3 and 300). Figure 6b confirms theoretical prediction[26] that there exists an optimal error rate performance for a finite number of training vectors n. This effect attenuates when the number of training vectors increase and disappears for n>6 (Figure 6c) where performance improves monotonically with the number of features when n=300 (Supplementary Figure S3). We can speculate at this stage that: (i) the propose concept requires variability. One or two OECTs are not enough to perform classification at a better level than chance. Since the output of each OECT is weighted by the perceptron (i.e. corresponding to synaptic weighting, signal summation and activation function), 12 OECTs without variability and providing the same response would be strictly equivalent to a perceptron with a single weight . This system, equivalent to a Boolean logic gate, cannot be used for classification task. (ii) The reservoir concept is rather resilient to the nature of variability presents in each individual OECT since absolute performance as a function of m is larger than mean deviation for each m values (see error bars in Figure 6b). In other words, the number of OECTs seems to play a more critical role on performances than the possible effect of a particular set of OECT (i.e. a particular set of features). The consequence of this point is to consider that increasing the number of OECTs allows coping with uncontrolled variability. More insights about the nature of variability should require higher OECTs' array with more complex input patterns in order to extract relevant trends between variability and performances.

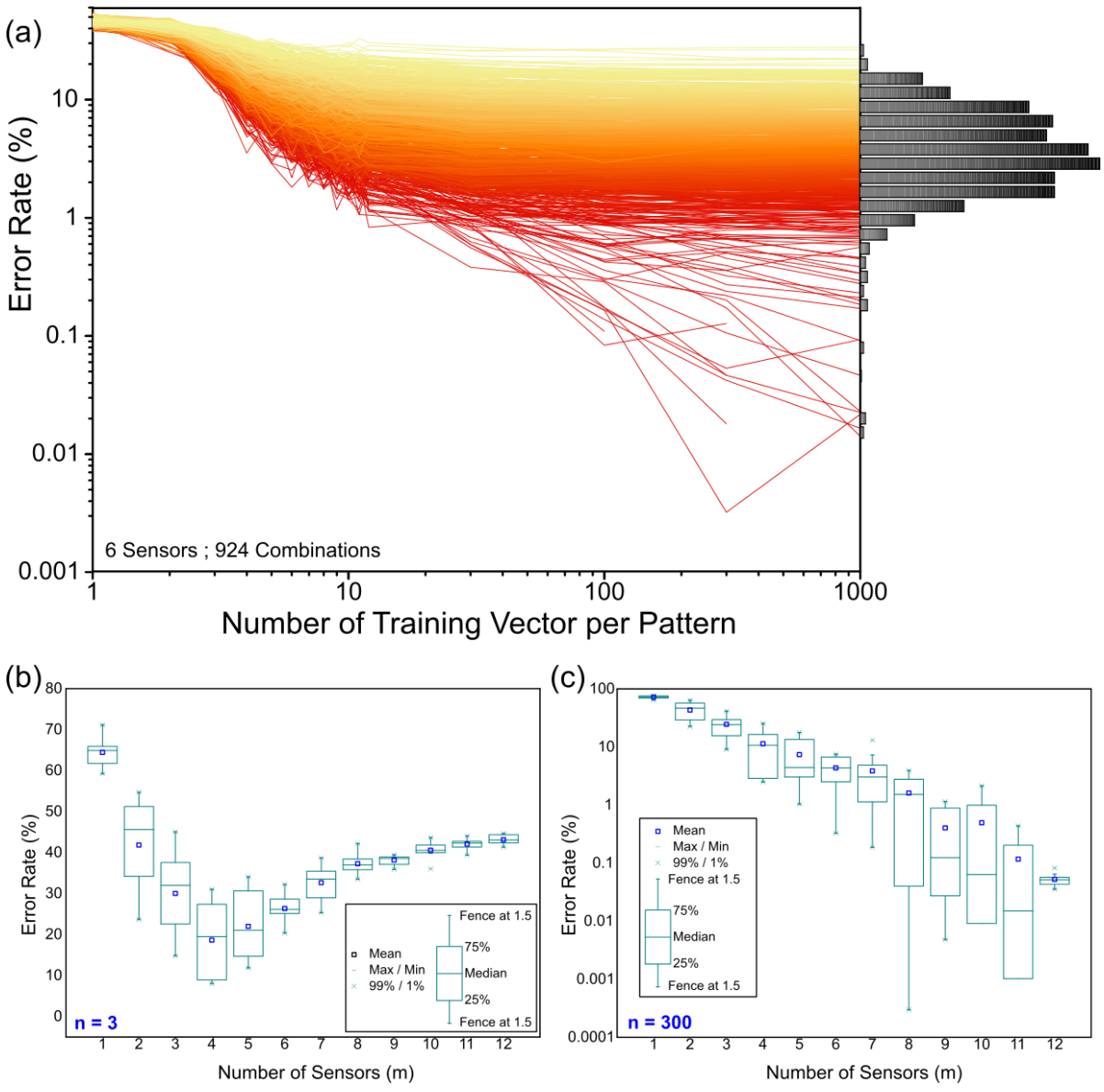

Figure 6. a, Error rate of the classification task obtained for all combination possible out of the 12 OECTs of the array as a function of number of training vectors. Performances for different array size m with b n=3 (b)

and n=300 (c) training vectors. Average performances as a function of m is obtained by calculating the mean value and deviation on 10 randomly chosen set of vectors for each m value (vector set displayed in Supplementary Table S1).

4.3. CONCLUSION

We demonstrated in this study that despite the high level of inherent variability in our bottom-up ion-sensing devices, we successfully discriminated dynamic patterns out of our OECT network by using a well-established neuromorphic learning algorithm. We showed that the RC approach can efficiently cope with many-fold variabilities in both transistor characteristic time constants and non-linear current levels to recognize frequency-modulated pulsed signal. Although it was not possible to correlate the dispersion of each device properties with the recognition performance of the OECT array, this work shows that neuromorphic sensing takes advantage of these variabilities, considered as drawbacks for standard sensing. While today's transistor technologies are based on high reproducibility and fast response, we point out the potential to rethink the excellence criteria for sensing in a neuromorphic data analysis context, and open up possible directions for future sensing technologies based on variability-rich materials, grown by bottom up approaches.

References

[1] 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, D. S. Modha, *Science* **2014**, *345*, 668. link
[2] W. Maass, T. Natschläger, H. Markram, *Neural Comput.* **2002**, *14*, 2531. link
[3] H. Jaeger, Technical Report GMD Report 148, German National Research Center for Information Technology, (**2001**). link
[4] M. Lukoševičius, H. Jaeger, *Comput. Sci. Rev.* **2009**, *3*, 127. link
[5] L. Appeltant, M. C. Soriano, G. van der Sande, J. Danckaert, S. Massar, J. Dambre, B. Schrauwen, C. R. Mirasso, I. Fischer, *Nat. Commun.* **2011**, *2*, 468. link
[6] Y. Paquot, F. Dupont, A. Smerieri, J. Dambre, B. Schrauwen, M. Haelterman, S. Massar, *Sci. Rep.* **2012**, *2*, 287. link
[7] R. Martinenghi, S. Rybalko, M. Jacquot, Y. K. Chembo, L. Larger, *Phys. Rev. Lett.* **2012**, *108*, 244101. link
[8] J. Torrejon, M. Riou, F. Abreu Araujo, S. Tsunegi, G. Khalsa, D. Querlioz, P. Bortolotti, V. Cros, K. Yakushiji, A. Fukushima, H. Kutoba, S. Yuasa, M. D. Stiles, J. Grollier, *Nature* **2017**, *547*, 428. link
[9] K. Vandoorne, P. Mechet, T. van Vaerenbergh, M. Fiers, G. Morthier, D. Verstraeten, B. Schrauwen, J. Dambre, P. Bienstman, *Nat. Commun.* **2014**, *5*, 3541. link
[10] F. Hempel, J. K.-Y. Law, T. C. Nguyen, W. Munief, X. Lu, V. Pachauri, A. Susloparova, W. T. Vu, S. Ingebrandt, *Biosens. Bioelectron.* **2017**, *93*, 132. link
[11] D. Khodagholy, T. Doublet, P. Quilichini, M. Gurfinkel, P. Leleux, A. Ghestem, E. Ismailova, T. Hervé, S. Sanaur, C. Bernard, G. G. Malliaras, *Nat Commun.* **2013**, *4*, 1575. link
[12] X. Gu, C. Yao, Y. Liu, I.-M. Hsing, *Adv. Healthcare Mater.* **2016**, *5*, 2345. link
[13] C. B. Nielsen, A. Giovannitti, D.-T. Sbircea, E. Bandiello, M. R. Niazi, D. A. Hanifi, M. Sessolo, A. Amassian, G. G. Malliaras, J. Rivnay, I. McCulloch, *J. Am. Chem. Soc.* **2016**, *138*, 10252. link
[14] A. Giovannitti, D.-T. Sbircea, S. Inal, C. B. Nielsen, E. Bandiello, D. A. Hanifi, M. Sessolo, G. G. Malliaras, I. McCulloch, J. Rivnay, *Proc. Natl. Acad. Sci.* **2016**, *113*, 12017. link
[15] J. T. Mabeck, J. A. DeFranco, D. A. Bernards, G. G. Malliaras, *Appl. Phys. Lett.* **2005**, *87*, 013503. link
[16] D. A. Bernards, G. G. Malliaras, *Adv. Funct. Mater.* **2007**, *17*, 3538. link
[17] P. Gkoupidenis, N. Schaefer, B. Garlan, G. G. Malliaras, *Adv. Mater.* **2015**, *27*, 7176. link
[18] P. Gkoupidenis, N. Schaefer, X. Strakosas, J. A. Fairfield, G. G. Malliaras, *Appl. Phys. Lett.* **2016**, *107*, 263302. link
[19] H. Markram, D. Pikus, A. Gupta, M. Tsodyks, *Neuropharmacology* **1998**, *37*, 489. link
[20] Y. van Burgt, E. Lubberman, E. J. Fuller, S. T. Keene, G. C. Faria, S. Agarwal, M. J. Marinella, A. A. Talin, A. Salleo, *Nat. Mater.* **2017**, *16*, 414. link
[21] R. Legenstein, W. Maass, What makes a dynamical system computationally powerful? New Directions in Statistical Signal Processing: From Systems to Brains 127–154 (MIT Press, **2007**). link
[22] P. G. Erlandsson, N. D. Robinson, *Electrophoresis* **2011**, *32*, 784. link
[23] J. L. Elman, *Cognit. Sci.* **1990**, *14*, 179. link

[24] J. Torrejon, M. Riou, F. Abreu Araujo, S. Tsunegi, G. Khalsa, D. Querlioz, P. Bortolotti, V. Cros, K. Yakushiji, A. Fukushima, H. Kutoba, S. Yuasa, M. D. Stiles, J. Grollier, arXiv:1701.07715 (v1). link
[25] B. Schrauwen, D. Verstraeten, J. Van Campenhout, In Proceedings of the 15th European Symposium on Artificial Neural Networks (ESANN'2007). 471-482 (**2007**). link
[26] G. Hughes, *IEEE Trans. Inf. Theory* **1968**, *14*, 55. link
[27] T.-K. Tran, Q. Bricaud, M. Oçafrain, P. Blanchard, J. Roncali, S. Lenfant, S. Godey, D. Vuillaume, D. Rondeau, *Chem. Eur. J.* **2011**, *17*, 5628. link
[28] D. Khodagholy, M. Gurfinkel, E. Stavrinidou, P. Leleux, T. Herve, S. Sanaur, G. G. Malliaras, *Appl. Phys. Lett.* **2011**, *16*, 163304. link
[29] S. Pecqueur, S. Lenfant, D. Guérin, F. Alibart, D. Vuillaume, *Sensors* **2017**, *17*, 570. link

5. CHAPTER 5

An iono-electronic neuromorphic interface for communication with living systems

5.1. INTRODUCTION

Our understanding of the computing principles originating our capacity to realize complex tasks, learn and adapt to our environment has done tremendous progresses in the last few decades. Boosted by technological breakthrough such as advanced imaging techniques with unprecedented resolution (i.e. two-photons imaging[1], for example) and computer performances for modeling large populations of cells[2], new perspectives to understand this complex machinery are foreseen. Sustained by large-scale projects such as the BRAIN initiative in US, HBP flagship in EU, the MIND project in Japan and more recently the China Brain Initiative, more is expected in the coming decades[3]. From the applicative side, this revolution brings lots of hope with the development of complex neuroprosthesis for arms and legs replacement[4] toward full exoskeleton, artificial sensors for vision[5] and audition[6] and brain disease treatment via deep brain stimulation[7] or basic brain degeneracy understanding.

Nevertheless, this appealing trend should be balance by fundamental limitations that are already appearing and that we need to address to sustain this evolution: if our understanding of the brain has made huge progresses, **we are still inefficient in interfacing biological systems with electronics, both in terms of energy and integration potential**. Pushed by the need to use conventional computers for building complex systems dedicated to brain interface applications, we have mostly capitalized on technologies and architectures inherits from microelectronic that are intrinsically not adapted to interface living systems. Standard electronics relies on fast charges (i.e. electrons) confined in (semi)conductors while biological neural networks capitalize extensively on distributed ions with slow dynamics to reach ultra-low power consumption and tolerance to noise and variations. **I propose to explore a new field of research by developing innovative materials and devices that would bridge ideally these two world.** In addition to rethinking the electronic used to interface the brain circuitry a real breakthrough would be to change our conception of data processing and signal representation. While conventional electronics separate physically sensing and computing, with the inherent fundamental bottleneck effect for data exchange, brain computing merges completely signal transmission and computing in its highly parallel organization of neural cells.

Reproducing bio-inspired concepts directly at the interface and designing devices and circuit matching this computing paradigm could profoundly improve our ability to interact with the brain.

I propose in the IONOS project to shift the brain interface paradigm by developing new technologies designed to interact intimately with biological cells and to bridge ideally the biological and the artificial world (figure 1). I will exploit an optimal computing paradigm, the bio-inspired computing (or neuromorphic computing) to implement an innovative interface to the living system. I will capitalize on **iono-electronic materials** that offer an **optimal transduction from ionic signal to electronic one (and reciprocally)** thanks to their mix ionic-electronic charges transport properties. **Designing devices** based on these materials and using ions dynamics for implementing innovative computing functions will offer the possibility to **replicate key biological computing features** at work in biological neural networks. **My approach clearly proposed to merge sensing and computing at the interface in order to optimize communication between the biological medium and an artificial system**. I propose to demonstrate efficient communication between an artificial system and a living system by integrating these technologies in a neuromorphic interface that will be connected to in-vitro cells' culture.

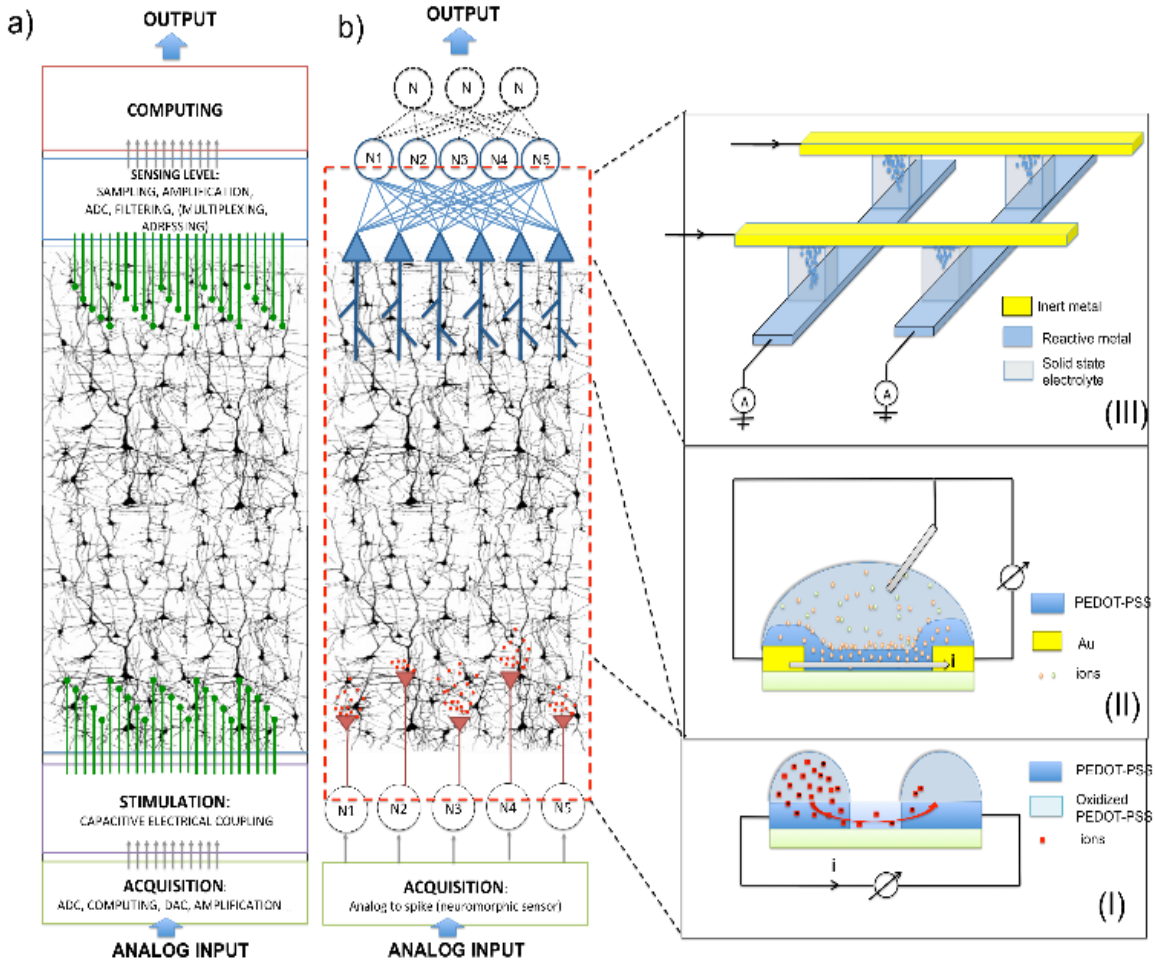

Figure 1: Communication consists in sending an analog input signal (a sensory stimulus, for example) to the neural cells and to record/compute the evoked activity in order to extract some meaningful output (an action, for example). (a) Conventional communication architecture used to interface biological neural networks. Electrical stimulation and sensing are performed with discrete electrodes (green). (b) The proposed concept of neuromorphic interface merging sensing and computing based on iono-electronic devices. Stimulation is realized via neurotransmitters (NT) release with (I) ionic pump (red triangle). Spatio-temporal integration and recording of the cells' activity is realized by (II) dendritic-like OECTs that extend into the neural cells assembly. The amplified signal by OECT is transmitted directly to the memory device implementing the (III) artificial synaptic array

5.2. OBJECTIVES: BRINGING NEUROMORPHIC ENGINEERING AT THE INTERFACE WITH BIOLOGY

While conventional approaches to brain interfaces consider separately the sensing and computing level, these two concepts are ubiquitous in biological networks. Computing features are embedded in the way information is exchanged between cells. When a spike is generated at the soma and propagates along the axono-dendritic tree to finally release neurotransmitters and stimulate the next cells, lots of elementary processes are contributing to the overall spike processing. When organized into large network, these elementary processes give rise to collective effects and to complex computing functions.

Neuromorphic engineering[8], which aims at developing material implementations of these bio-inspired computing features, is considered as a promising way for interfacing living systems since it capitalizes on the same computing paradigm and offers advance energy consumption performances for high-end computing tasks. Nevertheless, it has only been considered separately from the biological world and rely on standard interface technologies for connecting it. **Extending the neuromorphic concepts at the interface where the hardware interact with the biological world is my objective and represents the clear novelty of the IONOS project**. In addition, mainstream researches in sensing technologies rely on smart sensors with a rich set of conventional features (amplification, ADC, multiplexing).[9] **I propose here a new strategy of sensing based on bio-inspired features. This approach will unlock the fundamental bottleneck for exchanging information between two systems, the biological and the artificial one**.

5.2.1. Objective 1: in-situ synaptic learning on biological signals with resistive memory devices

It is now well recognized that biological networks acquired their functionality through synaptic plasticity. Spiking activity in the network modify the synaptic weight between two neurons, thus strengthening meaningful information paths and weakening non-useful one. **I will capitalize on the ability of resistive memory devices, or memristive devices, to mimic synapses to realize this function directly at the interface with living cells (figure 2)**. The basic idea consists in taking advantage of tunable analog resistive memory to implement in a dense and energy efficient array the synaptic weight[10]. Several works have successfully implemented variations around synaptic learning with resistive memory devices reproducing the Long Term Plasticity (LTP) observed in biological synapses during learning[11-15]. Going into the detail of the physics of resistive switching, direct analogies between ions dynamics in the memory cell and synaptic processes observed in biology have been established. Short Term Plasticity (STP), corresponding to a volatile memory effects, was implemented with organic transistors[16]. Capitalizing on the mix drift-diffusion of ions in resistive memory, STP to LTP transition, reminiscent of memory consolidation in biology, was successfully demonstrated[17-19]. More recently, a direct analogy between Ca^{2+} delivery during spiking and ions dynamics in Electro Chemical Metallization (ECM) cells was proposed.[20]

Nevertheless, **these interesting features are today only considered as a solution for computing in artificial systems**. I propose to demonstrate that this ability of **memory devices to mimic biology can be used to interact directly with living cells and to realize the fundamental synaptic learning directly on biological signals**. The major limitation of this approach is the gap in switching voltage and current of memory elements and the effective voltage produced by a cell on a passive electrode (<1mV). **The challenge is then to develop further memory technologies with ultra-low switching current and voltages** in order to reduce this gap. Some of our preliminary work[18,21] and other from literature[22] have shown that ECM cells were promising solutions to reach this goal ($V < 100$ mV and $I < 1\mu A$). Other technologies based on 2D materials such as MoO_x/MoS_2 will be also considered thanks to their promising performances demonstrated recently[23,24]. Since $V < 1$mV is unlikely to switch any memory device, **we will capitalize on local amplification realized by Organic Electro Chemical Transistors (OECT)** that demonstrated record transconductance[25] well adapted to drive a memory array and which are ideal sensors.

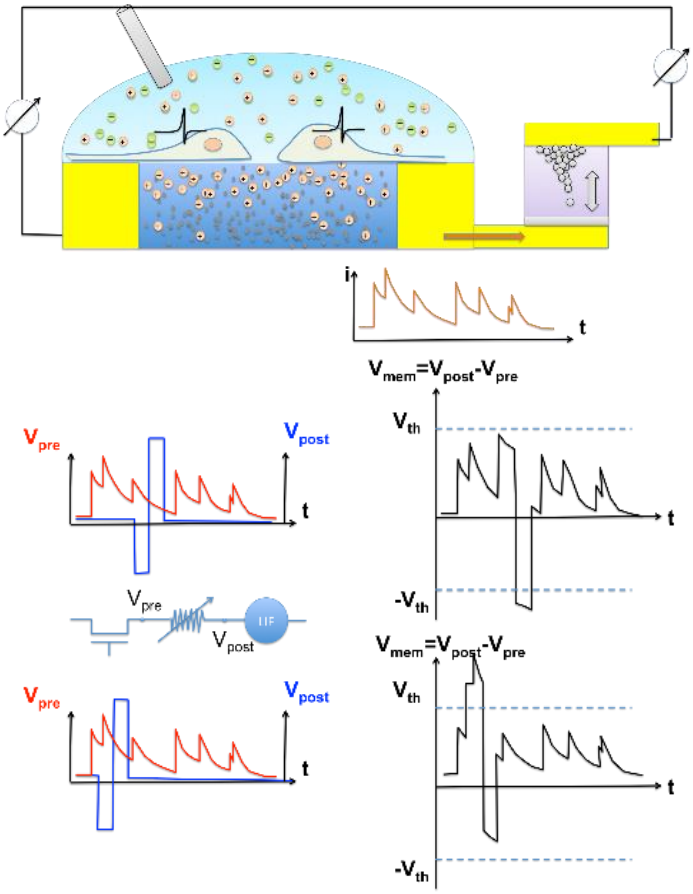

Figure 2 : (top) the OECT perform recording and integartion of the biological signal while memory device adapt and learn on the signal. (bottom) Example of learning on a biological signal with an artificial feedback. The memory potentiates (depresses) when V is above (below) the threshold

5.2.2. Objective 2: Spatio temporal integration of the signal with dendritic sensors and synaptic actuators.

Computation in biological networks is largely realized by successive integrations of the transmitted signal. Spatial extension of the axono-dendritic tree is very efficient for spatially integrating the neighboring cells' activity. Ion dynamics across the membrane and NT release at the synaptic terminal reproduce also a rich set of temporal integration processes. We propose to engineer on-purpose iono-electronic devices that will provide spatio-temporal integration in addition to optimal sensing and actuation. Realizing both computing and sensing based on ionic processes represents a challenge that we propose to address in this project.

Iono-electronic devices with mix ionic-electronic transport properties represent a shift in the basic working's principle of sensing/actuating since ions-electrons interaction is not based anymore on capacitive coupling across an interface (surface effect) but results from manipulation of ions penetrating the material (bulk effect) and interacting with electronic charges. At the sensing side, Organic Electro-Chemical Transistors[25] (OECT) where ionic signal is transduced into an electronic one via organic material (de)doping, offer local amplification (with high transconductance thanks to their bulk transport properties) and have demonstrated record performances for electrical activity recording in living systems[26]. At the stimulating side, since transistors are not adapted to electrical stimulation, the same basic iono-electronic materials have been used to implement Ionic Pump (IP)[27]. In this technology, the electronic signal is converted into a flux of ions. In particular, IP have demonstrated the ability to deliver neurotransmitters (NT) such as Acetylcholine to in-vitro living cells stimulation[28].

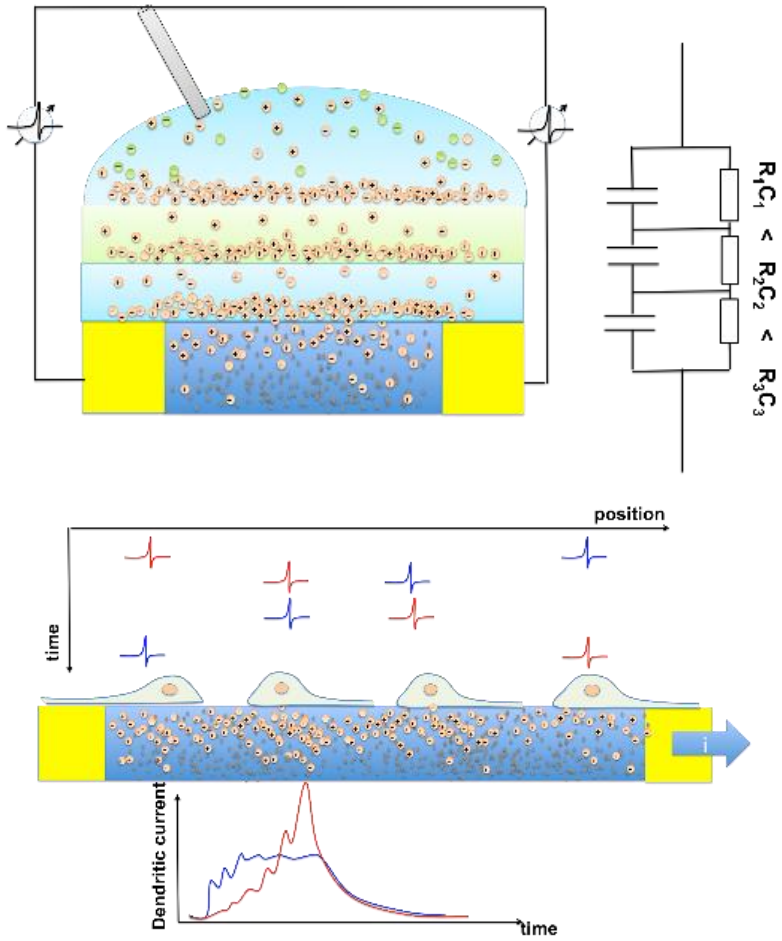

Figure 3 : (top) concept of metaplasticity implemented with multi-layer stacking with different permeability to ions. (bottom) concept of dendritic computing. Output current is a combination of spatial and temporal integration showing preferential direction

I propose to capitalize on these technologies and on the dynamics of ions exchange in these solid-state devices to reproduce spatio-temporal integration features observed in biology for sensing and stimulating (figure 3). Ions exchange across the PEDOT/liquid medium interface is likely to reproduce temporal integration observed in biological networks. Using ions dynamics, STP effects have been reported in OECT[29]. I will push further this idea by engineering the permeability of PEDOT to ions in order to tune the effective capacitance of the organic materials and consequently adapting the temporal response of the sensor. **This strategy project to implement the metaplasticty concept** proposed in computational neurosciences[30] as an optimal memorization mechanism in biological networks. Spatial integration will be realized by engineering long OECT that will interact with multiple cells. Recent works have shown that localization of stimulation on OECT can induce spatial signal integration[31]. This effect will be used for **implementing dendritic computing functions[32].** At the stimulation side, IP will be further developed to **reproduce the dynamics of NT release at the synaptic connection such as synaptic facilitation and synaptic depression.**

5.2.3. Objective 3: demonstration of efficient communication on a classification task.

In order to demonstrate the capability of my interface, I target to demonstrate a bio-inspired function: pattern classification. **Efficient communication between a biological system and an artificial one will be demonstrated on a neuromorphic platform that will integrate the different technologies developed previously.** Proof of concept will be realized on in-vitro neural cells' network coupled to the neuromorphic interface (figure 4).

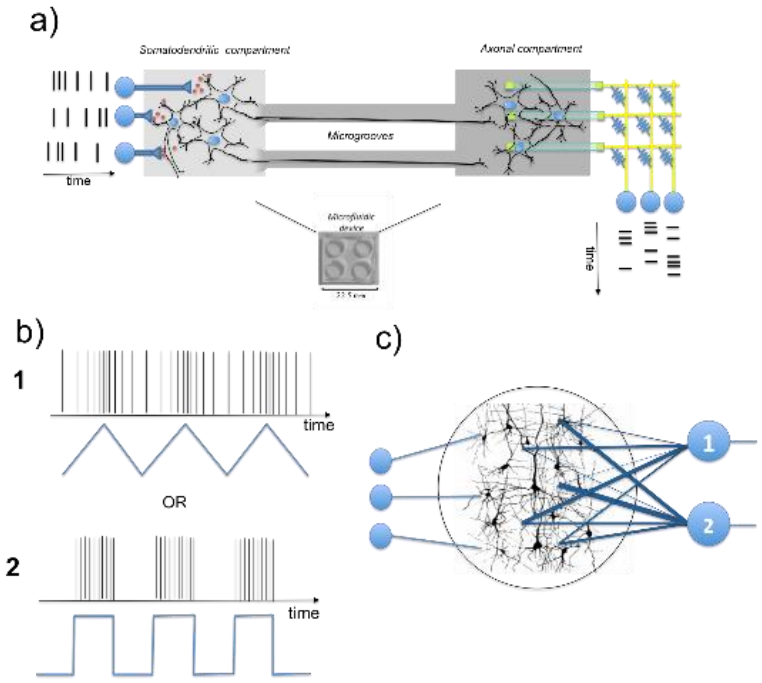

Figure 4 (a) physical implementation. IP stimulates cells with NT release. OECT sense the cell activity and amplify current for the memory array. (b) example of two analog signal to be classified and converted in spike. (c) Schematic of RC. Output neurons activates in response to signal 1 or 2 after proper learning.

By communication, I mean a global system that can send, receive and process information for some meaningful task. I will demonstrate that a biological network associated to the neuromorphic interface can realize the basic function of classification. I will implement a computing task inspired by the recent work around the concept of reservoir computing (RC)[33]. RC has been simultaneously developed for machine learning (echo state network[33]) and computational neurosciences (Liquid state machine[34]) and will provide a well-adapted formalism to understand and explore the overall computing system capability.

RC is based on the idea of projecting an input signal on a large and random recurrent neural network of high dimension. The projected signal is sampled with a limited read-out circuitry (a basic single layer perceptron) that performs classification through learning (adaptation of the weights of the read-out layer). RC capitalizes on both the reservoir non-linearity and on learning for realizing complex tasks. In terms of structure, RC topology is directly transposable to our concept: the input will be realized by the ionic pump stimulation, the reservoir will be associated to the biological cells network and the read-out circuitry will be implemented by the dendritic OECT and adaptive synaptic array.

An example of the classification task is presented in figure 4. Conventional supervised learning and bio-inspired learning of the read-out synaptic array with STDP or other learning/adaptation mechanisms developed during the project will be implemented on the platform. **This proof of concept will clearly demonstrate our ability to communicate with biological network and will open new avenues for brain computing deciphering and for brain machine interactions.**

5.3. BREAKTHROUGH, IMPACT AND COMPLEMENTARITY WITH OTHER APPROACHES

The breakthrough of this project is to propose a **new interface concept between artificial systems and living cells**. This approach is possible by adopting an innovative strategy at the material and device engineering level. I will develop a **new class of electronic** that will extensively capitalize on ionic dynamics for implementing features closer to biological world such as synaptic learning and spatio-temporal integration. This idea open new directions for device engineering, with new functions and concepts extracted from

computational neurosciences and biology. Iono-electronic materials are only in their infancy. Our project will challenge these materials with new engineering objectives and practical implementations.

Bringing new materials and devices in the neuromorphic computing field will increase the perspectives for neuromorphic engineers. Developing practical solutions for dendritic computing and advanced synaptic plasticity (i.e. metaplasticity) represent **a new direction for neuromorphic engineering.** Furthermore**,** capitalizing on fully analog and time dependent signal through spatio-temporal integration represents a **real shift in the way we conceive signal representation and processing** in brain machine interfaces.

My approach also proposes a **shift in the sensing paradigm**. I propose to develop neuromorphic concepts at the interface and consequently to **merge sensing and computing** at this level. This approach is reminiscent of the recent field of "compressed sensing". Reproducing the intrinsic communication principles observed in biological networks directly at the interface offers a direct solution to the bottleneck effect between sensing and computing.

This project targets a very strategic position since the interface is a natural bridge between different scientific communities. The neuromorphic interface will stimulate new interactions between biology and computer science. My approach will also directly benefit from/to recent research efforts at the computing level such as the ERCs NeuroAgents or NanoInfer, the EU RAMP or Brainbow project (i.e. brain-inspired computing). Development in neuroengineering with the Connexio ERC project developing in-vitro cell's cultures and in biology in general with organ-on-chip and 3D cell's culture will also strongly benefit to our long-term perspectives of interfacing more complex biological systems. I have already developed strong relationship with various leaders in these neighboring fields to ensure strong scientific exchanges and research stimulation.

References :

1. SVOBODA, Karel et YASUDA, Ryohei. Principles of two-photon excitation microscopy and its applications to neuroscience. *Neuron*, vol. 50, no 6, p. 823-839 (2006)
2. ANANTHANARAYANAN, Rajagopal, ESSER, Steven K., SIMON, Horst D., *et al.* The cat is out of the bag: cortical simulations with 109 neurons, 1013 synapses. In : *High Performance Computing Networking, Storage and Analysis, Proceedings of the Conference on*. IEEE. p. 1-12. (2009)
3. GRILLNER, Sten, IP, Nancy, KOCH, Christof, *et al.* Worldwide initiatives to advance brain research. *Nature neuroscience*, vol. 19, no 9, p. 1118-1122 (2016)
4. HOCHBERG, Leigh R., SERRUYA, Mijail D., FRIEHS, Gerhard M., *et al.* Neuronal ensemble control of prosthetic devices by a human with tetraplegia. *Nature*, vol. 442, no 7099, p. 164-171 (2006)
5. www.2-**sight**.com
6. www.**cochlear**.com
7. BERGMAN, Hagai, WICHMANN, Thomas, et DELONG, Mahlon R. Reversal of experimental parkinsonism by lesions of the subthalamic nucleus. *Science*, vol. 249, no 4975, p. 1436-1438 (1990)
8. MEAD, Carver. Neuromorphic electronic systems. *Proceedings of the IEEE*, vol. 78, no 10, p. 1629-1636 (1990)
9. HA, Sohmyung, AKININ, Abraham, PARK, Jiwoong, *et al.* Silicon-Integrated High-Density Electrocortical Interfaces. *Proceedings of the IEEE*, vol. 105, no 1, p. 11-33 (2017)
10. STRUKOV, Dmitri B. Nanotechnology: smart connections. *Nature*, vol. 476, no 7361, p. 403-405 (2011)
11. ALIBART, Fabien, ZAMANIDOOST, Elham, et STRUKOV, Dmitri B. Pattern classification by memristive crossbar circuits using ex situ and in situ training. *Nature communications*, vol. 4 (2013)
12. PREZIOSO, Mirko, MERRIKH-BAYAT, Farnood, HOSKINS, B. D., *et al.* Training and operation of an integrated neuromorphic network based on metal-oxide memristors. *Nature*, vol. 521, no 7550, p. 61-64 (2015)
13. BURR, Geoffrey W., SHELBY, Robert M., SIDLER, Severin, *et al.* Experimental demonstration and tolerancing of a large-scale neural network (165 000 synapses) using phase-change memory as the synaptic weight element. *IEEE Transactions on Electron Devices*, vol. 62, no 11, p. 3498-3507 (2015)
14. JO, Sung Hyun, CHANG, Ting, EBONG, Idongesit, *et al.* Nanoscale memristor device as synapse in neuromorphic systems. *Nano letters*, vol. 10, no 4, p. 1297-1301 (2010)

15. KUZUM, Duygu, JEYASINGH, Rakesh GD, LEE, Byoungil, *et al.* Nanoelectronic programmable synapses based on phase change materials for brain-inspired computing. *Nano letters*, vol. 12, no 5, p. 2179-2186 (2011)
16. ALIBART, Fabien, PLEUTIN, Stéphane, GUÉRIN, David, *et al.* An organic nanoparticle transistor behaving as a biological spiking synapse. *Advanced Functional Materials*, vol. 20, no 2, p. 330-337 (2010)
17. DU, Chao, MA, Wen, CHANG, Ting, *et al.* Biorealistic implementation of synaptic functions with oxide memristors through internal ionic dynamics. *Advanced Functional Materials*, vol. 25, no 27, p. 4290-4299 (2015)
18. LA BARBERA, Selina, VUILLAUME, Dominique, et ALIBART, Fabien. Filamentary switching: Synaptic plasticity through device volatility. *ACS nano*, vol. 9, no 1, p. 941-949 (2015)
19. OHNO, Takeo, HASEGAWA, Tsuyoshi, TSURUOKA, Tohru, *et al.* Short-term plasticity and long-term potentiation mimicked in single inorganic synapses. *Nature materials*, vol. 10, no 8, p. 591-595 (2011)
20. WANG, Zhongrui, JOSHI, Saumil, SAVEL'EV, Sergey E., *et al.* Memristors with diffusive dynamics as synaptic emulators for neuromorphic computing. *Nature materials*, vol. 16, no 1, p. 101-108 (2017)
21. LA BARBERA, Selina, VINCENT, Adrien F., VUILLAUME, Dominique, *et al.* Interplay of multiple synaptic plasticity features in filamentary memristive devices for neuromorphic computing. *Scientific Reports*, vol. 6 (2016)
22. SUN, Xiangyu, WU, Chuangui, SHUAI, Yao, *et al.* Plasma-Induced Nonvolatile Resistive Switching with Extremely Low SET Voltage in TiO x F y with AgF Nanoparticles. *ACS Applied Materials & Interfaces*, vol. 8, no 48, p. 32956-32962 (2016)
23. BESSONOV, Alexander A., KIRIKOVA, Marina N., PETUKHOV, Dmitrii I., *et al.* Layered memristive and memcapacitive switches for printable electronics. *Nature materials*, vol. 14, no 2, p. 199-204 (2015)
24. CHENG, Peifu, SUN, Kai, et HU, Yun Hang. Memristive behavior and ideal memristor of 1T phase MoS2 nanosheets. *Nano letters*, vol. 16, no 1, p. 572-576 (2015)
25. KHODAGHOLY, Dion, RIVNAY, Jonathan, SESSOLO, Michele, *et al.* High transconductance organic electrochemical transistors. *Nature communications*, vol. 4 (2013)
26. KHODAGHOLY, Dion, DOUBLET, Thomas, QUILICHINI, Pascale, *et al.* In vivo recordings of brain activity using organic transistors. *Nature communications*, vol. 4, p. 1575 (2013)
27. ISAKSSON, Joakim, KJÄLL, Peter, NILSSON, David, *et al.* Electronic control of Ca2+ signalling in neuronal cells using an organic electronic ion pump. *Nature materials*, vol. 6, no 9, p. 673-679 (2007)
28. TYBRANDT, Klas, LARSSON, Karin C., KURUP, Sindhulakshmi, *et al.* Translating electronic currents to precise acetylcholine–induced neuronal signaling using an organic electrophoretic delivery device. *Advanced materials*, vol. 21, no 44, p. 4442-4446 (2009)
29. GKOUPIDENIS, Paschalis, SCHAEFER, Nathan, GARLAN, Benjamin, *et al.* Neuromorphic functions in PEDOT: PSS organic electrochemical transistors. *Advanced Materials*, vol. 27, no 44, p. 7176-7180 (2015)
30. FUSI, Stefano, DREW, Patrick J., et ABBOTT, Larry F. Cascade models of synaptically stored memories. *Neuron*, vol. 45, no 4, p. 599-611 (2005)
31. GKOUPIDENIS, Paschalis, KOUTSOURAS, Dimitrios A., LONJARET, Thomas, *et al.* Orientation selectivity in a multi-gated organic electrochemical transistor. *Scientific reports*, vol. 6 (2016)
32. LONDON, Michael et HÄUSSER, Michael. Dendritic computation. *Annu. Rev. Neurosci.*, vol. 28, p. 503-532 (2005)
33. LUKOŠEVIČIUS, Mantas et JAEGER, Herbert. Reservoir computing approaches to recurrent neural network training. *Computer Science Review*, vol. 3, no 3, p. 127-149 (2009)
34. BUONOMANO, Dean V. et MAASS, Wolfgang. State-dependent computations: spatiotemporal processing in cortical networks. *Nature Reviews Neuroscience*, vol. 10, no 2, p. 113-125 (2009)

ANNEXE: scientific resume

EDUCATION

2004-2008 PhD THESIS UPJV Amiens-(FRANCE)

Laboratoire de Physique des Couches Minces (LPMC), M. BENLAHSEN

Realization and characterization of Amorphous Carbon Thin Films (optoelectronics and microstructural properties).

Realization and development of carbon-based electronic devices (OLED and Organic FET).

CURRENT POSITION

2020 to present PERMANENT RESEARCHER IEMN-CNRS (FRANCE)

Permanent researcher at IEMN-CNRS. Université de Lille.

Development of memristive circuits and devices for neuromorphic computing. Front-end and Back-end interfaces for electrophysiology.

PREVIOUS POSITIONS

2017 - 2020 PERMANENT RESEARCHER LN2-3IT/CNRS (Sherbrooke - CANADA)

Associated to the LN2-3IT laboratory, University of Sherbrooke, as part of international exchange between CNRS and UdeS.

Development of memristive circuits and devices for neuromorphic computing. Development of BEOL integration of memristive devices. Exploration of new integration strategies for 3D memory devices architectures.

2012 to 2017 PERMANENT RESEARCHER IEMN-CNRS (FRANCE)

Rank 1st among 120+ candidates for the tenure position. PI of the neuromorphic devices team - Nanostructure, Components and Molecules group (NCM)

Development of memristive circuits and devices for neuromorphic computing. Fabrication and integration of oxide-based memory (TiO_2 and HfO_2), electrochemical metallization memory (Ag/Ag_2S) and organic-based memory. Opto-electronic interfaces to in-vitro cells' culture.

2010 to 2012 POST-DOC Univ. California Santa Barbara (USA)

Novel Electronic Devices and Computing System Group, D. B. STRUKOV

Starting the experimental activity of the group by developing the processes (cleanroom facility) and characterization setup (electrical measurement) of the memristive devices.

Development of analog memory devices, circuits and neuromorphic applications

2008-2010 POST-DOC IEMN-CNRS (FRANCE)

Nanostructure, Components and Molecules group (NCM), D. VUILLAUME

Associated to the European Project NABAB (ICT-FP7) - C. GAMRAT

Realization and characterization of neuro-inspired synaptic devices (NPs/OTFT technology)

Implementation of neuro-inspired functions.

SUPERVISION OF GRADUATE STUDENTS AND POSTDOCTORAL FELLOWS

Supervision of students (% of supervision / subject / source of funding):

- Dr. Gina Adams (2010-2012), PhD candidate at UCSB. Now assistant professor at University of Washington
  (20%, fabrication and electrical characterization of memristor)
- Dr. Selina La Barbera (2013-2016), PhD candidate at IEMN. Now associate editor at Nature Communications.
  (80%, engineering of synaptic plasticity with ECM devices)
- Marie Minvielle (2013-2017), PhD candidate at Institut of Nanotechnology of Lyon, France. Now lecturer at Université de Poitiers.
  (33%, engineering of TiO2 thin films through MBE for memristor engineering)
- Pierrick Charlier, PhD student, sept. 2017 to sept 2018. Quitted. Move to engineer position
  (33%, fabrication of memristor and co-integration on CMOS for machine learning accelerators, HIDATA/NSERC)
- Abdelouadoud El Messoudy, PhD student, 2018-2022, Universite de Sherbrooke.
  (25%, development of BEOL compatible fabrication process for TiO2 memristors heterogeneous integration. HIDATA/NSERC)
- Raphael Dawant, PhD student, sept. 2019, Universite de Sherbrooke
  (20%, fabrication of 3D crossbar arrays for ultra high density memristor integration, HIDATA/NSERC)
- Waqas Bashir, PhD student, sept. 2019, Universite de Sherbrooke. Quite in september 2020
  (33%, fabrication of memristor and co-integration on CMOS for machine learning accelerators, HIDATA/NSERC)
- Mahdi Ghazal, PhD student, sept. 2019, IEMN
  (33%, development of PEDOT:PSS sensors for electrophysiology, IONOS/ERC)
- Kamila Janzakova, PhD student, sept. 2019, IEMN
  (33%, neuromorphic dendritic devices based on electropolymerization of PEDOT materials, IONOS/ERC)
- Ismael Balafrej, PhD student, Jan. 2020, Université de Sherbrooke
  (50%, spiking neural network design and integration of hardware constraints into SNN models, UNICO/ChistEra)
- Corentin Scholaert, PhD student, oct. 2021, université de Lille
  (33%, dendritic materials and devices for neuromorphic sensing, IONOS/ERC and regionHdF)
- Nikhil Garg, PhD student, oct. 2021, cotutelle Univ. Lille / Univ. Sherbrooke
  (33%, CMOS/memristor hardware for neuromorphic computing, IONOS/ERC)
- Alexis Melot, PhD student, sept. 2021, univ. Sherbrooke
  (50%, sparse SNN design for bio-signals encoding and processing, Chaire Neuromorphic UdeS)
- Davide Florini, PhD student, jan. 2022, Univ. Sherbrooke

(25%, integration of memristors with CMOS for SNN with on-chip learning, UNICO/ChistEra)
- Benoit Manchon,, PhD student, oct. 2021, cotutelle UdeS / Univ. Lyon (20%, fabrication and characterzation of HZO ferroelectric memristors, bourse IRL-LN2/ univ. Lyon)
- Joao Quintino Palhares, PhD student, oct 2021, cotutelle UdeS / STMicroelectronics / UGA (20%, exploration of non-volatile memory technologies for SNNs implementation on 28 nm CMOS, CIFRE/STMicroelectronics)

Supervision of postdocs:
- Dr. Gilbert Sassine (2013-2016), postdoc at IEMN. Now associate researcher at CEA, France
- Dr. Nabil Najjari (2013-2016), posdoc at IEMN. Now IT engineer
- Dr. Anna Susloparova (2016 to 2020), postdoc at IEMN / JPArc. Interface nanotechnology / biology
- Dr. Sebastien Pecqueur (2018-2019), postdoc IEMN, now permanent researcher CNRS
- Dr. Ankush Kumar, (2019 to 2022), postdoc IEMN

GRANTS (with major implication)

2012-2015 **DINAMO** (PI) ANR – Retour postdoc
Development and integration of memory devices for neuromorphic computing (480k€)

2015-2016 **M2NP** (PI) PEPS-CNRS
Development of a multifunctionnal platform for interface with living systems
Coll. IEMN / JPArc / UdeMons (France) (40k€)

2017-2020 **HIDATA** (co-PI) Strategic-NSERC
Integration and packaging of passive crossbar arrays for machine learning applications
Coll. UdeS / UdeT (Canada) (1M$)

2018-2023 **IONOS** (PI) Discovery-NSERC
Development of iono-electronic materials for neuromorphic computing (110k$)

2018-2023 **IONOS** (PI) ERC
A neuromorphic interface with living systems based on iono-electronic materials
@ UdeS, IEMN, JPArc (1,9 M€)

2020-2023 **UNICO** (Co-PI) Chist-Era
Developpement of a spiking neural network toolkit for Edge Computing applications.
Coll. UdeS, C2N,UAM, IBM-Zurich, CEZAMAT-PW (1 M€)

2022-2027 **Chaire neuromorphic** (Co-PI) MEI/UdS
Neuromorphic computing and engineering SW/HW co-design.
Research chair with Pr. Sean Wood (1 M$)

ORGANISATION OF SCIENTIFIC MEETINGS

2014 : Scientific board member, 1st National Workshop of GdR Oxyfun, Autrans, France (70 participants)

2015 : Executive/Scientific board member, 1st National Workshop of GdR BioComp, St-Paul-de-Vence, France (70 participants)

2016 : Executive/Scientific board member, 2nd National Workshop of GdR BioComp, Lyon, France (50 participants)

2016 : Part of the Technical Program Commitee of NANOARCH 2016, Beijing, China.

2017 : Scientific board member of *Material and Device Integration on Silicon for Advanced Applications* symposium at EMRS fall meeting 2017, Warsaw, Poland.

INSTITUTIONAL RESPONSIBILITIES

2014 – 2017: co-responsible de l'axe nanoélectronique du laboratoire LN2 – Sherbrooke.
2014 – present: Founding member and executive board member of the BioComp GdR / French researcher network with interest in implementation of neuromorphic computing (about 150 participants)/France
2014 – 2017 : Executive board member of GdR Oxyfun, French researcher network with interest in functional oxides (about 200 participants)/France

COMMISSIONS OF TRUST

2016 - 2017 : Expert in the Neuromorphic group of the French Observatory of Micro and Nano Technologies (OMNT).
2012-present : Member of the Examining Committee of a PhD defence (UCSB, CNRS-Thales, CNRS-IEF, UPJV-LPMC)
2010-present: reviewer for Nature Communications, Scientific Reports, Sciences Advances, ACS Nano, ACS Materials and Interfaces, Advanced Materials, Advanced Functional Materials, Advanced Electronics Materials, Applied Physics Letters, IEEE TED, IEEE Nanotechnology, IEEE TCAS, Journal of Applied Physics.

SHORT / MID TERM VISIT

March 2018-July 2018: visiting scientist in D. Strukov's group. UCSB

MAJOR COLLABORATIONS

Damien Querlioz, France. We collaborate on device modeling and high level computing strategies exploration. I am currently expert of his ERC NanoInfer for device fabrication.
Dmitri Strukov: UCSB/ECE department (US). We collaborate on implementations of neuromorphic computing based on memristive devices from out two labs.
Guilhem Larrieu (LAAS), Timothée Levy (IMS): We have launched a collaborative initiative gathering neuro-engineering French experts for students exchange and expertise sharing. This initiative is sustained by CNRS-INSIS.
Dominique Drouin : UdeS, Chair at IBM-Bromont, Sherbrooke, Canada : We collaborate on integartion and packaging of memory devices.
Luc Buée : JPArc Laboratory, INSERM. I launched our collaboration with the M2NP PEPS-CNRS project in 2015 (role : PI). We develop MEA and microfluidic for in-vitro cells' recording

SCIENTIFIC PRODUCTION

Regular papers in peer-review journals

[48] P-CRITICAL: a reservoir autoregulation plasticity rule for neuromorphic hardware
I Balafrej, F Alibart, J Rouat
Neuromorphic Computing and Engineering 2 (2), 024007, 2022
[47] Theoretical modeling of dendrite growth from conductive wire electro-polymerization
A Kumar, K Janzakova, Y Coffinier, S Pecqueur, F Alibart
Scientific reports 12 (1), 1-11, 2022
[46] Exploiting non-idealities of resistive switching memories for efficient machine learning
V Yon, A Amirsoleimani, F Alibart, RG Melko, D Drouin, Y Beilliard
Frontiers in Electronics, 8, 2022
[45] CODEX: Stochastic Encoding Method to Relax Resistive Crossbar Accelerator Design Requirements
T Liu, A Amirsoleimani, J Xu, F Alibart, Y Beilliard, S Ecoffey, D Drouin, R Genov
IEEE Transactions on Circuits and Systems II: Express Briefs, 2022
[44] Bio-Inspired Adaptive Sensing through Electropolymerization of Organic Electrochemical Transistors
M Ghazal, M Daher Mansour, C Scholaert, T Dargent, Y Coffinier, S Pecqueur, F Alibart
Advanced Electronic Materials 8 (3), 2100891, 2022
[43] Fully CMOS-compatible passive TiO2-based memristor crossbars for in-memory computing
A El Mesoudy, G Lamri, R Dawant, J Arias-Zapata, P Gliech, Y Beilliard, S Ecoffey, A Ruediger, F Alibart, D Drouin

Microelectronic Engineering, 111706, 2022
[42] Dendritic organic electrochemical transistors grown by electropolymerization for 3D neuromorphic engineering
K Janzakova, M Ghazal, A Kumar, Y Coffinier, S Pecqueur, F Alibart
Advanced Science, 2102973, 2021
[41] Analog programing of conducting-polymer dendritic interconnections and control of their morphology
K Janzakova, A Kumar, M Ghazal, A Susloparova, Y Coffinier, F Alibart and S. Pecqueur
Nature communications 12 (1), 1-11, 2021
[40] Oxygen vacancy engineering of TaO x-based resistive memories by Zr doping for improved variability and synaptic behavior
JHQ Palhares, Y Beilliard, F Alibart, E Bonturim, DZ de Florio, FC Fonseca, D Drouin and AS Ferlauto
Nanotechnology 32 (40), 405202, 2021
[39] Multi-terminal memristive devices enabling tunable synaptic plasticity in neuromorphic hardware: a mini-review
Y Beilliard, F Alibart
Frontiers in Nanotechnology, 87, 2021
[38] Low impedance and highly transparent microelectrode arrays (MEA) for in vitro neuron electrical activity probing Auteurs
Anna Susloparova, Sophie Halliez, Séverine Begard, Morvane Colin, Luc Buée, Sébastien Pecqueur, **Fabien Alibart**, Vincent Thomy, Steve Arscott, Emiliano Pallecchi, Yannick Coffinier
Sensors and Actuators B: Chemical, 327-128895 (2021)
[37] In-Memory Vector-Matrix Multiplication in Monolithic Complementary Metal–Oxide–Semiconductor-Memristor Integrated Circuits: Design Choices, Challenges, and Perspectives
Amirali Amirsoleimani, **Fabien Alibart,** Victor Yon, Jianxiong Xu, M Reza Pazhouhandeh, Serge Ecoffey, Yann Beilliard, Roman Genov, Dominique Drouin
Advanced Intelligent Systems, 2, 11-2000115 (2020)
[36] AIDX: Adaptive Inference Scheme to Mitigate State-Drift in Memristive VMM Accelerators
Tony Liu, Amirali Amirsoleimani, **Fabien Alibart**, Serge Ecoffey, Dominique Drouin, Roman Genov
IEEE Transactions on Circuits and Systems II: Express Briefs (2020)
[35] Conductive filament evolution dynamics revealed by cryogenic (1.5 K) multilevel switching of CMOS-compatible Al2O3/TiO2 resistive memories
Yann Beilliard, François Paquette, Frédéric Brousseau, Serge Ecoffey, **Fabien Alibart**, Dominique Drouin
Nanotechnology, 31, 44-445205 (2020)
[34] Investigation of resistive switching and transport mechanisms of Al2O3/TiO2−x memristors under cryogenic conditions (1.5 K)
Yann Beilliard, François Paquette, Frédéric Brousseau, Serge Ecoffey, **Fabien Alibart**, Dominique Drouin
AIP Advances, 10, 2-025305 (2020)
[33] Physical mechanisms involved in the formation and operation of memory devices based on a monolayer of gold nanoparticles-polythiophene hybrid materials
T Zhang, D Guérin, **F Alibart**, D Troadec, D Hourlier, G Patriarche, ...
Nanoscale Advances (2019)
[32] A Compact Device Model for Nanoparticle-organic Memory Transistor's Characterization
H Van Mai, O Bichler, C Gamrat, Y Viero, F Alibart, D Vuillaume
Communications in Physics 28 (3), 191, (2018)
[31] Light-stimulatable molecules/nanoparticles networks for switchable logical functions and reservoir computing
Y Viero, D Guérin, A Vladyka, **F Alibart**, S Lenfant, M Calame, D Vuillaume
Advanced Functional Materials 28 (39), 1801506 (2018)
[30] Perspective: Organic electronic materials and devices for neuromorphic engineering
S Pecqueur, D Vuillaume, **F Alibart**
Journal of Applied Physics 124 (15), 151902 (2018)
[29] Neuromorphic Time-Dependent Pattern Classification with Organic Electrochemical Transistor Arrays
S. Pecqueur, …and **F. Alibart**,
Advanced Electronic Materials, accepted (2018)
[28] Cation discrimination in organic electrochemical transistors by dual frequency sensing
S Pecqueur, D Guérin, D Vuillaume, **F Alibart**
Organic Electronics 57, 232-238 (2018)
[27] Electron-transport polymeric gold nanoparticles memory device, artificial synapse for neuromorphic applications
B. Hafsi et al.,
organic Elecctronics 50, 499-506 (2017)
[26] Negative Differential Resistance, Memory, and Reconfigurable Logic Functions Based on Monolayer Devices Derived from Gold Nanoparticles Functionalized with Electropolymerizable TEDOT Units

T Zhang, D Guerin, **F Alibart**, et al.,
The Journal of Physical Chemistry C, 121,18 (2017)
[25] Concentric-electrode organic electrochemical transistors: case study for selective hydrazine sensing
S Pecqueur, S Lenfant, D Guérin, **F Alibart**, D Vuillaume
Sensors 17 (3), 570 (2017)
[24] Interplay of multiple synaptic plasticity features in filamentary memristive devices for neuromorphic computing
La Barbera, S., Vincent, A. F., Vuillaume, D., Querlioz, D., & **Alibart, F**.
Scientific Reports, *6*. (2016)
[23] Neuromorphic computing based on emerging memory technologies
B. Rajendran**, F. Alibart**
IEEE JETCAS. 6, 2 198-211 (2016)
[22] Interfacial versus filamentary resistive switching in TiO2 and HfO2 devices,
G. Sassine, S. La Barbera, N. Najjari, M. Minvielle, C. Dubourdieu, **F. Alibart**
JVST-B, 34, 012202 (2016)
[21] Low voltage and time constant organic synapse-transistor,
S. Desbief, A. Kyndiah, D. Guerin, D. Gentili, M. Murgia, S. Lenfant, **F. Alibart**, T. Cramer, F. Biscarini, D. Vuillaume
Organic Electonics, 21, 47-53 (2015)
[20] Modeling and experimental demonstration of a Hopfield Network Analog-to-Digital converter with hybrid CMOS/Memristor circuits
X. Guo, F. Merrikh-Bayat, L. Gao, B. D Hoskins, **F. Alibart**, B. Linares-Barranco, L. Theogarajan, C. Teuscher, D. B Strukov
Frontiers in Neurosciences, 9 (2015)
[19] Filamentary switching: synaptic plasticity through device volatility
S. La Barbera, D. Vuillaume and **F. Alibart**,
ACS Nano, 9, 941-949 (2015)
[18] Plasticity in memristive devices for spiking neural networks
S. Saïghi, C. G Mayr, T. Serrano-Gotarredona, H. Schmidt, G. Lecerf, J. Tomas, J. Grollier, S. Boyn, A. F Vincent, D. Querlioz, S. La Barbera, **F. Alibart**, D. Vuillaume, O. Bichler, C. Gamrat, B. Linares-Barranco,
Frontiers in neurosciences, 9 (2015)
[17] Pattern Classification by Memristive Crossbar Circuits with Ex-situ and In-situ Training
F. Alibart, E. Zamanidoost, D. Strukov,
Nature Communication, 4, 2072 (2013)
[16] Programmable CMOS/memristor threshold logic
L. Gao, **F. Alibart**, D. Strukov,
IEEE Trans. Nanotechnology, 12, 115-119 (2013)
[15] Utilizing NDR effect to reduce switching threshold variations in memristive devices
F. Alibart , D. Strukov,
Applied Physics A, 111, 199-202 (2013)
[14] Pavlov's Dog Associative Learning Demonstrated on Synaptic-like Organic Transistors
O. Bichler, W. Zhao, **F. Alibart**, S. Pleutin, S. Lenfant, D. Vuillaume, Christian Gamrat,
Neural Computation, 25(2), 549-566 (2013)
[13] Ionically-mediated electromechanical hysteresis in transition metal oxides,
Y. Kim, A.N. Morozovska, A. Kumar, S. Jesse, E.A. Eliseev, **F. Alibart**, D. Strukov, S. V Kalinin,
ACS Nano 6 (8), pp. 7026-7033, (2012)
[12] Thermophoresis as the plausible mechanism for unipolar resistive switching in metal-oxide-metal memristors,
D.B. Strukov, **F. Alibart**, S. R. Williams,
Applied Physics A, 107, 509-518 (2012)
[11] High-precision tuning of state for memristive devices by adaptable variation-tolerant algorithm
F. Alibart, L. Gao, B. D. Hoskins, D. B. Strukov,
Nanotechnology IOP, 23, 075201 (2012)
[10] A memristive nanoparticle/organic hybrid synapstor for neuro-inspired computing,
F. Alibart, S. Pleutin, O. Bichler, C. Gamrat, T. Serrano-Gotarredona, B. Linares-Barranco, D. Vuillaume,
Advanced Funct. materials, 22, 609-616 (2012)
[7] Effect of nitrogen on the optoelectronic properties of a highly sp2-rich amorphous carbon nitride films
F. Alibart, M Lejeune, K Zellama, M Benlahsen.,
Diamond and Related Materials, Volume 20, Issue 3, Pages 409-412 (2011)
[6] The effect of the terminating bonds on the electronic properties of sputtered carbon nitride thin films
F. Alibart, S Peponas, S Charvet, M Benlahsen,
Thin Solid Films, 519, 3430-3436 (2011)

[9] Functional Model of a Nanoparticle Organic Memory Transistor for Use as a Spiking Synapse,
O. Bichler, W. Zhao**, F. Alibart**, S. Pleutin, D. Vuillaume, C. Gamrat,
IEEE Trans. on Electon. Devices, 57 (11), 3115 (2010)
[8] An organic-nanoparticle transistor behaving as a biological synapse,
F. Alibart, S. Pleutin, D. Guérin,, C. Novembre, S. Lenfant, K. Lmimouni, C. Gamrat, D. Vuillaume
Advanced Functional materials, 20, 330-337 (2010)
[5] Influence of Disorder on Localization and Density of States in Amorphous Carbon Nitride Thin Films Systems Rich in PI-bonded Carbon Atoms,
F. Alibart, M Lejeune, O Durand Drouhin, K Zellama, M Benlahsen,
J. Appl. Phys., 108, 053504 (2010)
[4] Comparison and semiconductor properties of nitrogen doped carbon thin films grown by different techniques.
F. Alibart, O Durand Drouhin, M Benlahsen, S Muhl, S Elizabeth Rodil, E Camps, L Escobar-Alarcon
Applied Surface Science, 254, 5564-5568 (2008)
[3] Evolution of the opto-electronic properties of amorphous carbon films as a function of nitrogen incorporation,
F. Alibart, F Alibart, O Durand Drouhin, M Lejeune, M Benlahsen, SE Rodil, E Camps
Diamond and Related Materials, 17, 925-930 (2008)
[2] Relationship between the structure and the optical and electrical properties of reactively sputtered carbon nitride films,
F. Alibart, O Durand Drouhin, C Debiemme-Chouvy, M Benlahsen,
Solid State Communications, 145, 392-396 (2008)
[1] Covalent grafting of organic molecular chains on amorphous carbon surfaces,
S. Ababou-Girard, F. Solal, B. Fabre, **F. Alibart**, C. Godet,
Journal of Non-Crystalline Solids, 352, 2011-2014 (2006)

Conference papers with peer-review

[19] Insertion of an Ultrathin Interfacial Aluminum Layer for the Realization of a Ferroelectric Tunnel Junction
B Manchon, G Segantini, N Baboux, P Rojo Romeo, R Barhoumi, IC Infante, F Alibart, D. Drouin, B Vilquin, D Deleruyelle
physica status solidi (RRL)–Rapid Research Letters, proceeding iof EMRS, 2100585, 2021
[18] Signals to Spikes for Neuromorphic Regulated Reservoir Computing and EMG Hand Gesture Recognition
N Garg, I Balafrej, Y Beilliard, D Drouin, **F Alibart**, J Rouat
International Conference on Neuromorphic Systems 2021, 1-8
[17] Addressing Organic Electrochemical Transistors for Neurosensing and Neuromorphic Sensing
Mahdi Ghazal, Thomas Dargent, Sebastien Pecqueur, **Fabien Alibart**
2019 **IEEE SENSORS**
[16] Observation of Highly Nonlinear Resistive Switching of Al2O3/TiO2-x Memristors at Cryogenic Temperature (1.5 K)
Y Beilliard, F Paquette, F Brousseau, S Ecoffey, **F Alibart**, D Drouin
NMDC 2020
[15] Memristor device characterization by scanning microwave microscopy
G Sassine, N Najjari, N Defrance, C Haenssler, D Theron, **F Alibart**, K. Haddadi
Manipulation, Automation and Robotics at Small Scales (**MARSS**), 2017
[14] Exploiting the Short-term to Long-term Plasticity Transition in Memristive Nanodevice Learning Architectures
C. H. Bennett, S. La Barbera, A. F. Vincent, J.-O. Klein, **F. Alibart** and D. Querlioz,
IJCNN (2016)
[13] Short-term to long-term plasticity transition in filamentary switching for memory applications
S. La Barbera, A.F. Vincent, D. Vuillaume, D. Querlioz, **F. Alibart**
Memristive Systems (MEMRISYS) International Conference on, 1-2 (2015)
[12] Neuromorphic hybrid RRAM-CMOS RBM architecture
M. Suri, V. Parmar, A. Kumar, D. Querlioz, **F. Alibart**
15th Non-Volatile Memory Technology Symposium (NVMTS), 1-6 (2015) *Best poster award*
[11] OXRAM based ELM architecture for multi-class classification applications
M. Suri, V. Parmar, **F. Alibart,**
IJCNN (2015)
[10] Pattern classification and recognition with memristive circuits
F. Alibart and D. Strukov,
DATE 2014 (2014)
[9] A reconfigurable FIR filter with memristor based weights
F. Merrikh-Bayat, **F. Alibart**, L. Gao, D. Strukov

ISCAS 15 (2014)
[8] Atomic switch: synaptic functionalities and integration strategies.
S. La Barbera, D. Guérin, D. Vuillaume, **F. Alibart**.
17èmes Journées Nationales du Réseau Doctoral en Micro-Nanoélectronique*, JNRDM 2014*, (2014)
[7] Digital-to-analog and Analog-to-Digital conversion with metal oxide memristors for ultra-low power computing
L. Gao, F. Merrikh-Bayat, **F. Alibart**, D.B. Strukov
NANOARCH 2013, (2013) *Best paper award*
[6] A high resolution nonvolatile analog memory ionic devices
L. Gao, **F. Alibart**, and D.B. Strukov
Non-Volatile Memories Workshop, (2013).
[5] Analog-input analog-weight dot-product operation with Ag/a-Si/Pt memristive devices
L Gao, **F Alibart**, DB Strukov
VLSI and System-on-Chip, (VLSI-SoC), IEEE/IFIP 20th International (2012)
[4] Analog-input analog-weight dot-product operation with Ag/a-Si/Pt memristive devices
L. Gao, **F. Alibart**, and D.B. Strukov,
Proceedings of VLSI-SoC'12, (2012)
[3] Hybrid CMOS/Nanodevice Circuits for High Throughput Pattern Matching Applications
F. Alibart, T. Sherwood, and D. B. Strukov,
Conference on Adaptive Hardware and Systems (2011)
[2] Development of a functional model for the nanoparticle-organic memory transistor
O. Bichler, W.S. Zhao, C. Gamrat, F. Alibart, S. Pleutin, D. Vuillaume
Proceedings of 2010 IEEE International Symposium on Circuits and Systems (2010)
[1] Fast and compact simulation models for a variety of FET nano devices by the CMOS EKV equations.
Serrano-Gotarredona, T., Linares-Barranco, B., Agnus, G., Derycke, V., Bourgoin, J. P., Alibart, F., Vuillaume, D., J. Sohn, J. Bendall, M. E. Welland, Gamrat, C.
Nanotechnology, 2009. IEEE-NANO *2009. 9th IEEE Conference on* (pp. 691-694). IEEE. (2009)

Invited talks

[17] Neuromorphic computing: a bridge between ANNs and biology.
F. Alibart
CNANO Nord Ouest, journée des prix de these. 2022
[16] Neuromorphic computing with organic materials and devcies
F. Alibart
Colloque NECOTIS / UAM, 2019, Sherbrooke
[15] Merging bio-sensing and neuromorphic computing with organic electro chemical transistors.
F. Alibart
MRS fall 2020, Boston
[14] Neuromorphic computing: a bridge between ANNs and biology
F. Alibart
TMU Symposium on Translational Applications of Biomedical Engineering Technologies, 2019
[13] Memristive devices for neuromorphic computing
F. Alibart
JMC 2018, Grenoble
[12] Neuromorphic computing: towards dynamical computing systems
F. Alibart
Materials Research Society Spring Meeting, MRS spring (2018)
Neuromorphic computing with Memristive devices: From ANNs to bio-inspired computing
F. Alibart
ISSNE 2017, Bordeaux
[11] Neuromorphic computing with memristive devices
F. Alibart
Journées EEA, Marseille (2017)
[10] Memristive devices: from bio-inspired computing to artificial neural networks.
F. Alibart
European Materials Research Society Fall Meeting, E-MRS Fall (2016)
[9] Développement de mémoires en architecture crossbar de haute densité pour un traitement de l'information neuromorphique
F. Alibart

7ème Colloque du Laboratoire Nanotechnologies Nanosystèmes, LN2 2016, Estrimont, QC, Canada, 10-13 juillet (2016)
[8] Memristive devices for neuromorphic computing.
F. Alibart
1er colloque national du GdR BioComp. St Paul de Vence, France (2015)
[7] RRAM devices : state of the art and challenges.
F. Alibart
1er colloque national du GdR BioComp. St Paul de Vence, France (2015)
[6] Neuromorphic computing with memory devices.
F. Alibart
1er colloque national du GdR Oxyfun. Autrans, France (2014)
[5] RRAM devices : state of the art and challenges.
F. Alibart
1er colloque national du GdR Oxyfun. Autrans, France (2014)
[4] Pattern classification with memristive crossbar arrays,
F. Alibart, E. Zamanidoost , D.B. Strukov
Nano and Giga Challenges, Phoenix, US (2014)
[3] Memristives technologies for neuromorphic applications
F. Alibart
Colloque mémoires non-volatiles émergentes, Nantes, France (2013)
[2] Neuromorphic computing with memristive devices,
F. Alibart
Workshop on functional oxides for integration in micro and nano electronics, Autrans, France (2013)
[1] Neuromorphic computing with memristive devices
F. Alibart, E. Zamanidoost, D. B. Strukov
EPICO, Buenos Aires, Argentina (2012)

Papers	48
Book chapter	1
Conference papers	19
Invited talks (national)	8
Invited talks (International)	9
Citations	3423
H-factor	24
i-10	37