Nanodevice-based Novel Computing Paradigms and the Neuromorphic Approach

Weisheng Zhao, Damien Querlioz, Jacques-Olivier Klein, Djaafar Chabi, Claude Chappert

IEF, Univ. Paris-Sud, CNRS; Orsay, 91405, France
weisheng.zhao@u-psud.fr and damien.querlioz@u-psud.fr

***Abstract*— Deep submicron (<90nm) Integrated Circuits (IC) suffer from both high static and dynamic power consumption, which are caused respectively by the growing leakage currents and large capacitance bus traffic. Nanodevice based novel computing paradigms are currently under intense investigation to overcome these issues and build up the next generation ICs performing with higher power efficiency and operating performance. In this paper, an overview and current status of this field is first presented, and then we focus on the memristive nanodevices based neuromorphic approach, which is considered as one of the most promising computing paradigms for power reduction and process variation or defect tolerance.**

I. Introduction

Traditional digital signal processing approaches suffer from both high static and dynamic power consumption at deep submicron (<90nm) complementary metal oxide semi-conductor (CMOS) technology and beyond [1]. For instance, the dominant computing model of microprocessor based on Von-Neumann architecture consumes much more power (e.g. ~ 1pJ@22nm node) to access the memory for fetching the instructions and reading/writing the data, than that of logic operation (e.g. ~ 1fJ at the 22nm node) [2-3] (See Fig.1). As the silicon memories for computing (e.g. SRAM) are intrinsically volatile, data in "idle" state should be always kept under the power supply and this leads to high static power, which is growing exponentially when scaling down the device features and begins to dominate the whole power dissipation of digital IC. In order to overcome power *bottleneck* and build up the next generation IC performing ultra low power while keeping high performance, research efforts on emerging nanodevices based novel computing paradigms has been started in early 2000's [4-5]. They promise to complement or replace respectively the CMOS technology and traditional computing approaches like multi-core and reconfigurable architectures. These fields get more and more attention after the rapidly rising of CMOS power wall and some significant technological advance of nanodevices [6-9].

This paper firstly overviews the emerging nanodevices currently under considerable investigation such as spintronics [6], memristors [9], graphene-based transistors [7] and multiferroic devices [10]. Their novel properties for energy saving and innovative computer architecture design are particularly addressed. In the following, diverse novel computing architectures promising ultra low power such as bio-inspired neuromorphic circuit [11], quantum cellular automata (QCA) [12] and material implementation logic (IMP) [13] and so on are presented, and are compared from the IC design point of view. At last a particular emphasis is drawn on the memristive nanodevice based neuromorphic approach, which promises excellent variation tolerance, ultra low power operation and high area efficiency etc.

This work is partially sponsored by the French national projects ANR-PANINI, NANOINNOV-SPIN, ANR-MARS and the CNRS PEPS projects: NV-CPU and Synapses.

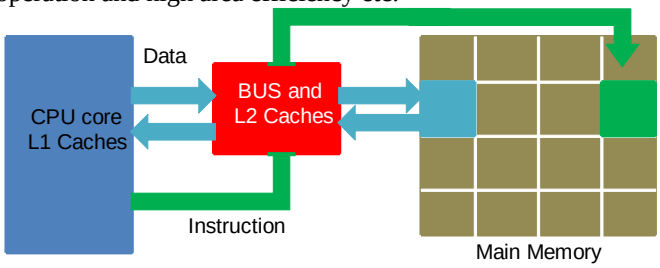

Figure 1. Volatile cache memories used for data access accleration leads to high static power. Data access between CPU core and main memory consumes much higher power than that of MOSFET transistor switching.

II. Emerging Nano-devices for Computing Systems

Nanodevices like nanowires, quantum dots, memristors, spintronics and graphene are intensely explored for different applications thanks to their advantageous characteristics beyond classical silicon devices [6-13]. For power saving purposes, the non-volatility and ultra low leakage currents are often mentioned as the most outstanding performances. For computing speed enhancement, higher electron mobility is the most important characteristics to be addressed. Nanodevices can be thus classified into the following three categories based on the main research interests.

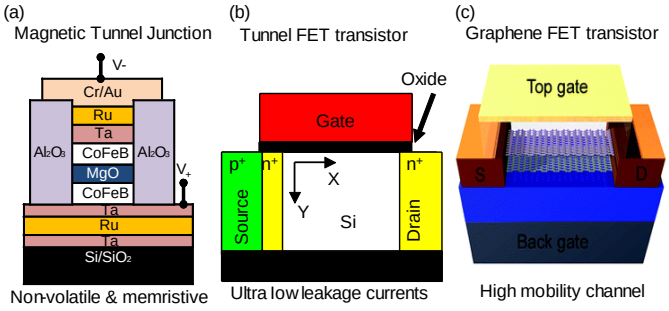

Figure 2. (a) a non-volatile and memristive nanodevice: magnetic tunnel junction composed of ferromagnetic and oxide thin films [6] (b) an ultra low I_{off} nanodevice: Si- channel based tunnel FET [14] (c) a nanodevice with high mobility channel: Graphene FEF transistor [7].

A. *Non-volatile and memristive devices*

In the next generation IC, non-volatile and memristive nanodevices are expected to be used as the data storage to provide eventual *dark silicon* [15]. The static power can be particularly reduced as the part of chip in "idle" state can be powered completely off [16]. A number of nanodevices have been demonstrated to present non-volatility and tunable resistance such as memristors, spintronics and multiferroic devices [6, 8-10]. These nanodevices attract special interest from semiconductor industries as they might be compatible

with standard CMOS processes and help relaxing the power wall for future feature size scaling [17]. From the technological point of view, these devices promise high maturity and low cost compared with post-silicon nano-electronics. In the last year, a number of pre-industrial demonstrators have been demonstrated. However all of them present lower speed and larger die area than silicon memories. For instance, magnetic tunnel junction (MTJ) is one of the most fast non-volatile and memristive devices (see Fig.2a), but its switching duration is physically limited to some nanosecond [6]. Novel computing paradigms are then necessary to integrate these devices and to obtain the good tradeoff between low power and high performance at the system level.

B. *Low leakage devices*

As mentioned above, traditional silicon transistors suffer from the growing leakage currents, which are dominated by I_{off}. Efforts to minimize leakage currents include the use of high-κ dielectrics, strained silicon, stronger doping levels and new device structures [1]. Si nanowire tunnel field-effect transistor (TFET) (see Fig.2b) is for example considered as a promising device to respond to the ultra-low leakage challenge of future feature sizes [14]. The benefits of the TFET are particularly linked to its potential of low sub-threshold slope (sub-60mV/dec) for V_{dd} scaling (see Eq.1 [18]). Furthermore, TFET could offer good system level energy efficiency for applications up to 1 GHz [14]. However, like other post-silicon nanodevices, TEFT undergoes important disadvantages like low I_{on} and low technological maturity.

$$E_b = N \times f_{op} \times \int_0^T V_{dd} \times I_d(t)dt \qquad (1)$$

C. *High Mobility devices*

For some years, the operating frequency of computing systems f_{op} has been saturating to some GHz even though the feature size scaling continues. This is due to the limited energy budget per chip E_b and exponential increasing of device number N (see Eq.1). One solution is to increase the electron mobility and then reduce the operating duration T. Graphene-based FET (GFET) transistors (see Fig.2c) became the preferred domain of condensed-matter and electron-device physicists in the last years thanks to the excellent mobility of graphene channel, which could be in excess of 15,000 $cm^2V^{-1}s^{-1}$, more than 10 times that of silicon [7]. They are considered as a promising option for post-silicon electronics [7, 17]. However, the nanofabrication of GFET based on bottom-up approaches leads to intrinsic high variation and defect rates at the wafer level. As conventional computing architectures are often sensitive to device defect and process variation, novel computing paradigms are required to integrate these nanodevices to perform reliable computing.

III. NOVEL COMPUTING PARADIGMS

Power reduction, area efficiency and speed enhancement to continue Moore's law will not only require new device and material solutions, but also proper system design based on novel architectures. As mentioned in the last section, nanodevices need also novel computing systems to benefit at most of their advantages and to overcome their shortcomings. With the current trends, the novel computing paradigms can be also classified into two following categories.

A. *Low power, small footprint or high speed*

As discussed in the previous section, novel computing paradigms are necessary to integrate non-volatile (NV) memristive nanodevices to lower power and reduce the die area. Some computing architectures presented in the past are now revisited thanks to the fast progress of NV memristive nanodevices beyond classical approach [19]. For instance, the architecture called "*logic in memory*" or "In memory processors" distributes the non-volatile memory cell in each computing operators (see Fig.3a) and removes the complex memory hierarchy for higher area efficiency [20-22]. In the last years, a number of new circuits based on "logic in memory" have been experimentally demonstrated [20-21]. As the memory cells are integrated in the local operators, the part of IC in "idle" state can be powered off and restarted instantaneously. This function allows zero standby power and a non-volatile CPU (*NV-CPU*) can be expected. The tighter integration of memory and CPU would shorten greatly the data access time and reduce the dynamic power (see Eq.1).

More revolutionary computing paradigms for higher power and area efficiency are also widely explored. For instance, in the multiferroic quantum cellular automata (MQCA) [12], each device shows a finite number of states with memristive resistance at a discrete time. The state of each cell is determined by the state of its adjacent cells at the last discrete time (see Fig.3b). Earlier, similar circuits had been proposed with coupled quantum dots. Functional cells have been demonstrated, but they are limited to extremely low temperature (typically in the dozens of mK range) [23]. Another example is material implication (IMP) logic (see Fig.3c) and the circuits have been recently demonstrated that use two or three elements of a one-dimensional memristor array [13].

The main challenge of these approaches is the important performance degradation with regards to traditional CMOS, which limits their use for wide applications.

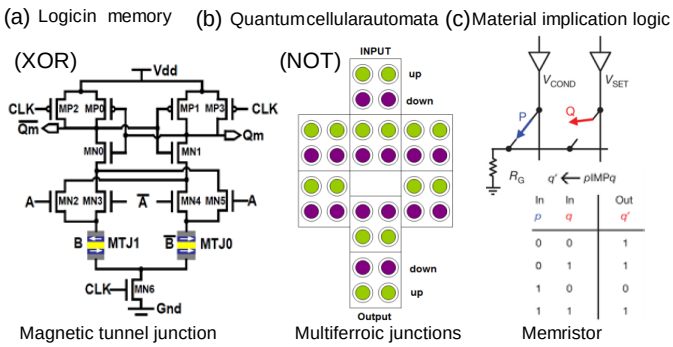

Figure 3. (a) Logic in memory architecture: XOR logic composed of magnetic tunnel junction memory cell [21] (b) Multiferroic cell based QCA NOT logic [12] (c) Memristor based IMP logic [13].

B. *High defect and variation tolerance*

Nanoscale devices present high intrinsic defect rate and important characteristics variation caused mainly by the nano-fabrication methods such as self-assembly and nano-imprint lithography [7-10]. Their circuit and system-level integration with high defect and variation tolerance becomes essential for practical applications. Self-adaptive, redundancy and error correction circuits have been intensely investigated to tolerate the defect and variation of nanodevices. Bio-inspired neural network is considered as one of the most promising approaches as it presents excellent defect and variation tolerance naturally

and also high performance in terms of power and speed benefiting from the massive parallelism [24-29]. In the next section, we will focus on this computing paradigm and show its good defect and variation tolerance through supervised and unsupervised learning.

IV. Memristive device based Neuromorphic Circuits

A. Motivation

A general difficulty for circuits based on nanodevices is the variability and low yield that most nanotechnologies possess. Biology can be an original inspiration to deal with this issue. Many researchers have stressed that the brain, for example, relies on variable and unpredictable neurons and synapses [25] and still manages a computational efficiency that outperforms our electronic systems. It is thus natural to wonder if we could imitate part of its essence to exploit our nanodevices. This question leads to novel computational paradigms and to a rethinking of how to use nanodevices. It has led to a rich literature [26–29], and is particularly explored for memristive devices. Being resistors that adjust their resistance depending on which voltage is applied to them, they are indeed reminiscent of synapses (the connections between neurons in the brain), which suggests they could be used as such. Additionally, the potentiality to integrate them as crossbars offers promises of extremely high integration, and thus of massive connectivity. In this paper, we describe two major strategies that are currently being explored into that direction, and how they deal with the variability issue.

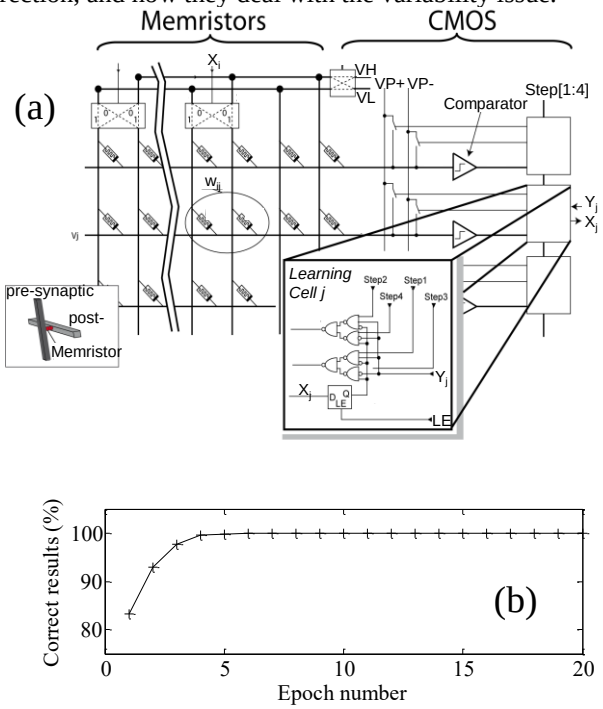

Figure 4. (a) Architecture of a "neural learning block" with a memristor crossbar array and CMOS neuron circuits. (b) Simulation of the system trained to learn the 104 linearly-separable digital functions with 3 inputs and 1 output. Proportion of correct answer of the network for successive presentation of a truth table ("epoch"). After 10 presentations, all the functions have been trained.

B. Supervised neuromorphic networks: nano-circuits that can be trained

One first lead is to envision the memristive device-based circuit as a reconfigurable unit cell that will not be programmed but *trained*, following ideas coming from the neural network field. It is indeed relatively straightforward to develop architectures where memristive devices behave as synapses of conventional, state-based artificial neural networks [28, 30]. Logical functions can then be trained to the circuit, in the same way an artificial network is trained with a dataset (see Fig.4). The benefit with regards to a traditional (lookup table-based) reconfigurable computing scheme is a natural high robustness to variability and defects [31]. Additionally, it can be further improved by a learning strategy that avoids defects – competitive learning – without the actual need to identify them. It has been shown that this approach increases robustness while requiring limited overhead in comparison with, for example, traditional error correction schemes [30].

The basic algorithm behind this approach already had experimental demonstration on the small scale with carbon-nanotube based devices [28] (the neural network learning rule was demonstrated). More works are needed to demonstrate real systems.

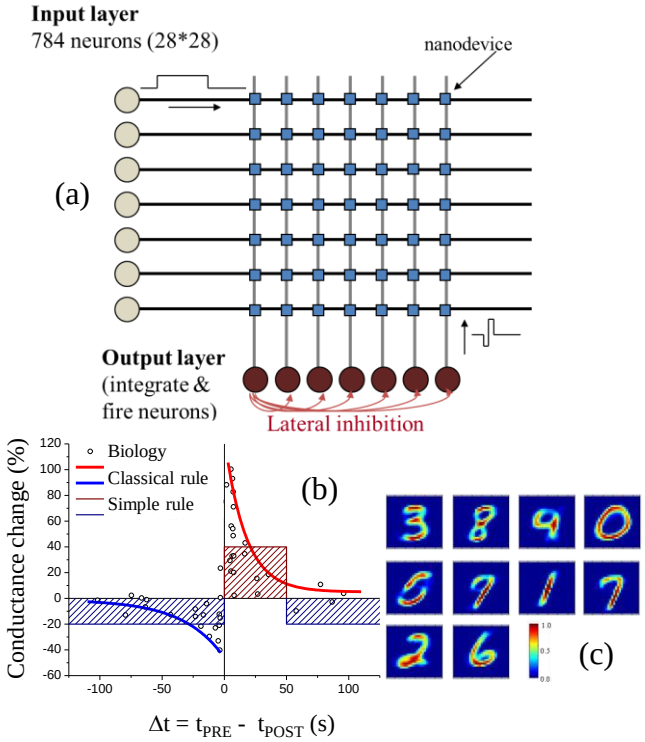

Figure 5. (a) Architecture of a spiking neural network for unsupervised learning. Circles are CMOS neurons. Squares, are memrsitive devices. The waveforms represent voltages pulses that implement STDP (b) The classical STDP rule and the simplified rule used in this architecture (c) Simulation results: a representation of devcies conductances after learning that shows handwritten digits' categorization.

C. Unsupervised neuromorphic networks: nano-circuit that can infer

An even more radical approach is to develop models of calculation that do not rely on traditional logic, escape usual paradigms, and are naturally immune to device variability. Unsupervised neural networks constitute a lead of particular interest. Such networks are not programmed as usual systems, but are instead able to infer regularities in data presented to them, and can perform cognitive-type functions. A vision to achieve that is the imitation of actual biological synapses. It has been suggested [32-33], and shown experimentally [34–36] that memristive devices could reproduce Spike Timing Dependent Plasticity (STDP), a behavior of synapses of the

brain [37]. STDP is valuable for synapses in spiking neural networks (where neurons communicate by emitting spikes or action potentials). Such networks can be implemented with asynchronous CMOS [38]. STDP synapses are sensitive to events when their pre- and post-synaptic neurons spike at close instants in time. If the pre-synaptic neuron spikes first, the synapse increases its conductance, if the postsynaptic neuron spikes first, the synapse decreases its conductance.

The possibility to implement STDP with memristive devices drives a lot of hope, even if it still raises questions. STDP has mostly been explored in computational neuroscience; few works have tried to use it for applications [39]. However, research is now being pursued in that direction. We present one example (see Fig.5). We use a highly simplified STDP rule, optimized to be implemented with memristive devices [40]. System-level simulations have shown that a system based on this rule can learn to categorize handwritten digits in an unsupervised way (see Fig.5). A striking feature of these simulations is that a device variability near-immunity is observed [40]. With standard variations as high as 50% of the mean value on all memristive device parameters, performance is barely affected. With variations of 100%, the performance decreases but the system is still functional. Indeed, due to the unsupervised nature of learning, sub-circuit learn features that they are naturally fit to learn due to device variability. In this context; variability is thus no longer a challenging issue to be solved, but a full part of the circuit functionality, which is the essence of this approach.

This kind of network has also been shown to be able of performing car counting on a video [41]. We hope that by scaling it to more devices and more complicated architectures, extremely complex tasks could be achieved with limited energy budget. We think that this approach is especially fit to categorize natural data, and could in particular provide smart low power sensors for embedded systems. On the longer term, such approaches open the possibility of cognitive computing, computer with real intelligence, a goal now pursued by several visionary projects [29], [42].

V. Conclusions and Perspectives

In this paper, we reviewed firstly the new nanodevices and novel computing paradigms under intense investigation. According to different research interests, they are classified and compared from the IC design point of view. Secondly, neuromorphic approach was particularly discussed to tolerate the high defect and variation rate of non-volatile memristive nanodevices through supervised or unsupervised learning.

References

[1] N.S. Kim et al., “leakage current: Moore's law meets the static power” IEEE Computer Society, pp.68-74, 2003.

[2] J. Backus, “Can programming be liberated from the von Neumann style?: a functional style and its algebra of programs”, ACM Turing award lectures, Doi:10.1145/1283920.1283933, 2007.

[3] M. Duranton, “New computing architectures for Green ICT”, Chist-era conference, 2011.

[4] D. B. Strukov and K. K. Likharev, “CMOL FPGA: a reconfigurable architecture for hybrid digital circuits with two-terminal nanodevices” Nanotechnol., vol. 16, no. 6, pp. 888-900, Jun. 2005.

[5] A. DeHon, “Array-based architecture for FET-based, nanoscale electronics,” IEEE Trans. Nanotechnol., vol. 2, no. 1, pp. 23-32, 2003.

[6] C. Chappert, A. Fert and F. Nguyen Van Dau, “The emergence of spin electronics in data storage” Nat. Mat., Vol.6, pp.813-823, 2007.

[7] F. Schwierz, “Graphene transistors”, Nat. Nanotech., Vol.5, 487, 2010.

[8] G. Agnus et al.,"2-Terminal Carbon Nanotube Programmable Devices for Adaptive Architectures "Advanced Material, Vol. 22, pp. 702-706, 2010.

[9] D. B. Strukov, G. S. Snider, D. R. Stewart, and R. S. Williams, “The missing memristor found,” Nature, vol. 453, no. 7191, pp. 80-83, 2008.

[10] J. Wang et al., “Epitaxial BiFeO3 Multiferroic Thin Film Hetero-structures”, Science, vol.299, pp.1719-1722, 2003.

[11] W.S. Zhao et al., “Nanotube devices based crossbar architecture: Toward neuromorphic computing”, Nanotechnology, Vol.21, 175202, 2010.

[12] M. Kabir et al., “RAMA: a self-assembled multiferroic magnetic QCA for low power systems”, in Proc. of GLSVLSI, pp.25-30, 2011.

[13] J. Borghetti, et al., “`Memristive’ switches enable `stateful’ logic operations via material implication,” Nature, vol. 464, pp. 873-876, 2010.

[14] K. Boucart et al., “Double-Gate Tunnel FET with High-K gate dielectric” IEEE Transactions on Electron Devices, vol. 54, pp. 1725-1733, 2007.

[15] N. Goulding et al., “The GreenDroid Mobile Application Processor: An Architecture for Silicon's Dark Future”, IEEE Micro Magazine, 86, 2011.

[16] W.S. Zhao et al., "Spin Transfer Torque (STT)-MRAM based Run Time Reconfiguration FPGA circuit" ACM Transactions on Embedded Computing Systems, Vol.9, No.2, article 14, 2009.

[17] International Technology Roadmap for Semiconductors, 2010.

[18] N.H.E. Weste and D.M. Harris, “CMOS VLSI Design: A Circuits and Systems Perspective” 4th Edition. Pearson/Addison-Wesley, 2010.

[19] W.S. Zhao et al., "New non-volatile logic based on spin-MTJ", Physica Status Solidi A, vol.6, pp. 1373-1377, 2008.

[20] S. Matsunaga et al., "Fabrication of a Nonvolatile Full Adder Based on Logic-in-Memory Architecture Using Magnetic Tunnel Junctions", Appl. Phys. Express (APEX), vol. 1, no. 9, pp. 091301-1~091301-3, Aug. 2008.

[21] W. Robinett et al., “A memristor-based nonvolatile latch circuit”, Nanotechnology, vol.21, 235203, 2010.

[22] Y. Gang, et al., “A High Reliability, Low Power Magnetic Full Adder”, IEEE Transactions on Magnetics, vol.47, pp.4611-4616, 2011.

[23] G. Snider et al., “Quantum-dot cellular automata: Review and Recent experiments”, J. Appl. Phys., vol.85, pp.4283-4285, 1999.

[24] G. A. Carpenter and S. Grossberg, “A massively parallel architecture for a self-organizing neural pattern recognition machine”, Computer Vision, Graphics and image processing, vol.37, pp.54-115, 1987.

[25] E. Marder et al., “Variability, compensation and homeostasis in neuron and network function,” *Nat. Rev. Neurosci.*, vol. 7, pp. 563-574, 2006.

[26] F. Alibart et al., “An Organic Nanoparticle Transistor Behaving as a Biological Spiking Synapse” Adv. Funct. Mater., vol. 20, 330, 2010.

[27] J. H. Lee and K. K. Likharev, “Defect-tolerant nanoelectronic pattern classifiers” Int. J. Circuit Theory Appl., vol. 35, no. 3, pp. 239-264, 2007.

[28] S.Y. Liao et al., “Design and Modeling of a Neuro-Inspired Learning Circuit Using Nanotube-Based Memory Devices,” IEEE Trans. Circuits Syst. Regul. Pap., vol.58, pp.2172-2181, 2011.

[29] G. Snider et al., “From Synapses to Circuitry: Using Memristive Memory to Explore the Electronic Brain” Computer, vol. 44, pp. 21-28, 2011.

[30] D. Chabi and J.-O. Klein, “High fault tolerance in neural crossbar” in Proc. of 2010 DTIS, pp. 1-6.

[31] D. Chabi, W.S. Zhao, D. Querlioz, and J.-O. Klein, “Robust neural logic block (NLB) based on memristor crossbar array,” in Proc. of 2011 NANOARCH, pp.137-143.

[32] G. S. Snider, “Spike-timing-dependent learning in memristive nanodevices,” in Prof. of NANOARCH, pp. 85-92, 2008.

[33] J. A. Pérez-Carrasco, et al., “On neuromorphic spiking architectures for asynchronous STDP memristive systems,” in Proc. of 2010 IEEE ISCAS, pp. 1659-1662.

[34] S. H. Jo et al., “Programmable Resistance Switching in Nanoscale Two-Terminal Devices,” Nano Lett., vol. 9, pp. 496-500, 2009.

[35] D. Kuzum et al., “Nanoelectronic Programmable Synapses Based on Phase Change Materials for Brain-Inspired Computing” Nano Letters, DOI: 10.1021/nl201040y, 2011

[36] S. Yu, et al., “An Electronic Synapse Device Based on Metal Oxide Resistive Switching Memory for Neuromorphic Computation,” IEEE Transactions on Electron Devices, vol. 58, pp. 2729-2737, Aug. 2011.

[37] G.-Q. Bi et al., “Synaptic modification by correlated activity: Hebb’s Postulate Revisited,” Annu. Rev. Neurosci., vol. 24, pp. 139-166, 2001.

[38] J. V. Arthur et al., “Learning in silicon: Timing is everything,” Advances in neural information processing systems, vol. 18, pp. 281-1185, 2006.

[39] T. Masquelier and S. J. Thorpe, “Unsupervised Learning of Visual Features through Spike Timing Dependent Plasticity” PLoS Comput Biol, vol. 3, no. 2, p. e31, Feb. 2007.

[40] D. Querlioz, et al., “Simulation of a memristor-based spiking neural network immune to device variations” Proc. of IJCNN, 1775, 2011.

[41] O. Bichler et al., “Unsupervised Features Extraction from Asynchronous Silicon Retina through Spike-Timing-Dependent Plasticity” Proc of IJCNN, pp. 859-866, 2011.

[42] D. S. Modha et al., “Cognitive computing” Communications of the ACM, vol. 54, p. 62, Aug. 2011.