ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

A Hybrid CMOS/Memristive Nanoelectronic Circuit for Programming Synaptic Weights

Arne Heittmann and Tobias G. Noll

Chair of Electrical Engineering and Computer Systems
RWTH Aachen University D-52062 Aachen, Germany

Abstract. In this paper a hybrid circuit is presented which comprises nanoelectronic resistive switches based on the electrochemical memory effect (ECM) as well as devices from a standard 40nm-CMOS process. A closed ECM device model, which is based on device physics, was used for simulations allowing for a precise prediction of the expected I-V characteristics. The device is used as a non-volatile and/or programmable synapse in a neuromorphic architecture. Expected performance figures are derived such as write time as well as robustness with regard to variations of supply voltage and timing errors. The results show that ECM cells are prospective devices for hybrid neuromorphic systems.

1 Introduction

Systems which are based on artificial neural networks (ANNs) are inherently parallel architectures. Since parallelism is one of the most powerful design principles to realize energy efficient systems while keeping performance high, ANNs are attractive candidates to be used in cognitive applications where energy is a limited resource and real time capabilities are mandatory. However, parallelism comes at the price of space. In neuromorphic architecture synapses provide the substantial amount of computing resources [1]. Since synapses are generally dynamic elements, i.e. they have to be programmed or adapted in order to optimize system performance, a particular challenge is given by providing adaptive elements with low space requirements. For most networks, the synaptic strength has to be adaptive with respect to a learning law or with a special dynamics modeling short term plasticity. In either case, storage of the concurrent synaptic strength is required. If multilevel strengths are necessary, multiple bits in memory (for a digital implementation) or analog memories such as capacitors (for analog implementations) are traditionally used which constitute an essential portion of the space demands for artificial synapses, see [1,2] for instance.

In this paper the focus is set on novel nanoelectronic devices, so-called resistive switches (RS) , which can be used as adaptive and/or non-volatile multi-level storage elements. RSs resemble the functionality of so-called memristors [3]. In principle, most types of resistive switches can be implemented with an area occupation of $4F^2$ (F: lithographic feature size) and hence outperform most established storage techniques with regard to area [4]. This is one reason why many researchers have proposed neuromorphic architectures comprising RSs as synapses [5,6,7,8,9]. However, resistive switches are passive devices, i.e. they cannot be used for signal amplification. Consequently, circuits of resistive switches have to be connected to an interface of active circuits in order to realize useful functions. If, for instance, scaled CMOS is brought together with RS devices, circuit designers are faced with the problem that voltages as well as current levels, which are necessary for robust RS operation, have to

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

match with those provided by standard CMOS devices. Although electrical prospects for CMOS are known there is still a considerable gap for the electrical requirements between RS device technology and CMOS, see [3] or [10] for typical RS device parameters. In addition, the design of hybrid RS-CMOS circuits is often hindered due to the absence of reliable physical models which can be used for circuit simulation. Typically, phenomenological device models are used instead [11]. This poses the risk of missing considerable device dynamics during simulation.

In this paper we will focus on a particular class of RS devices for which a closed physical device model has been developed recently [12]. The model was fully incorporated in our circuit simulation flow using SPICE allowing for the hybrid simulation of CMOS transistor devices and RS devices with representative parameters (see Fig. 4c or [12]). With that model at hand we were able to design circuits based on 40nm CMOS devices for programming the device resistance exploiting its multilevel capabilities as well as to derive performance figures. Hereby, the device is used as a simple synapse in a neuromorphic architecture. The results show that the electrical device properties match with the requirements of scaled CMOS. Hence, it could be used as a prospective device for future neuromorphic systems.

2 The ECM device model

So far, nine different electrically induced resistive switching effects have been identified [13]. The so-called ECM cells (**e**lectro**c**hemical **m**etallization) provide multilevel resistance programmability where resistances range from 10^{11} V/A down to 10^{6} V/A [13]. ECM devices are particularly interesting since the involved device currents can be hold in the nA-domain resulting in ultra-low power consumption during operation.

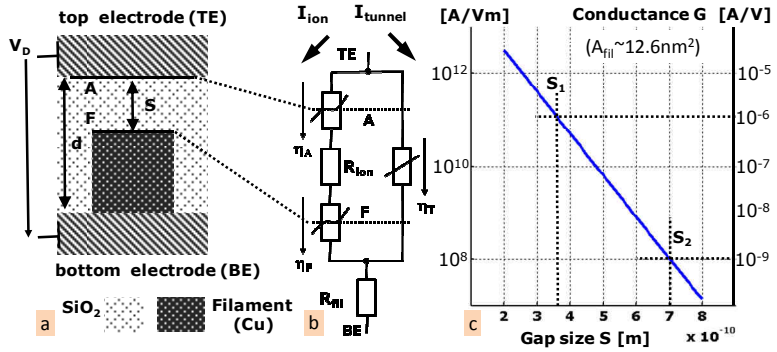

Fig. 1: a) ECM cell cross section; b) equivalent circuit; c) conductance vs. gap

In Fig. 1a a sketch of an ECM device based on Cu-SiO_2 is shown. The device consists of a bottom electrode as well as a top electrode which define the nodes for electronic access. Both electrodes are separated by an electronically insulating material (e.g. SiO_2 for Cu-SiO_2 cells) which also acts as an ion conducting layer. If a voltage V_D is applied, electrochemical active ions move into the ion conducting layer and drift towards the counterelectrode which is separated from the top electrode by a gap of size S. The deposition of ions on the counterelectrode results in the continuous growth of a so-called filament. The current density for the charge transfer across the electrolyte-

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

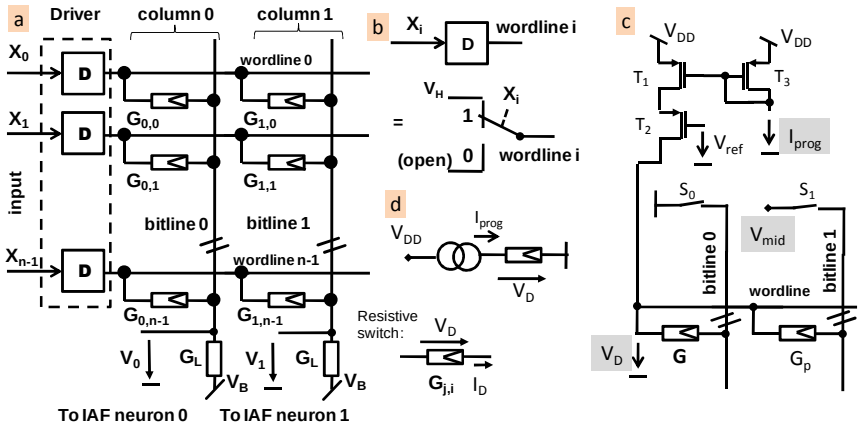

Fig. 2: a) synaptic array; b) wordline driver; c) programming setup, d) current src.

electrode interface during the cathodic reduction is described by the Butler-Volmer-equation [12]. As the ionic charge transport is associated with material transport towards the filament, the filament growth is in direct proportion to the ion current density and is reflected by a shrink in the gap size S:

$$\frac{dS}{dt} = -K_p \cdot 2 \cdot i_0 \sinh\left(\frac{z\eta_A}{2V_T}\right) = -K_p \cdot 2 \cdot i_0 \sinh\left(\frac{z\eta_F}{2V_T}\right) \qquad (1)$$

In (1) η_A (η_F) describes the voltage across the interface A (F), z is the number of charges per ion, V_T the temperature voltage and K_p as well as i_0 are specific constants which describe the relationship between matter deposition and electronic current density. For small gap sizes S the resistance R_{ion} can be neglected and the voltage across the interfaces is approximately half of the device voltage V_D. As S becomes significantly smaller than 1nm the electronic current becomes dominated by a tunneling current ($I_D=I_{ion}+I_{tunnel}$) which strongly depends on the gap size S [12]. For voltages V_D below 1V the overall device conductance established by I_{tunnel} can be approximated by (2) which constitutes a linear I-V relationship. Fig. 1c shows the conductance $G=I_D/V_D$ with regard to the gap size S. In (2) h, m_e, q, ϕ_0, A_{fil} denote the Planck constant, effective mass of an electron, elementary charge, barrier height and filament surface area. The conductance shows a variation over five orders of magnitude between S=0.3nm and S=0.8nm.

$$G \approx A_{fil} \cdot \frac{q^2}{4 \cdot \pi \cdot h \cdot S^2} \cdot e^{-S/L_0} \cdot \frac{S}{L_0} \qquad L_0 = \frac{1}{\sqrt{2m_e \cdot q \cdot \phi_0}} \cdot \frac{h}{4 \cdot \pi} \qquad (2)$$

Combining (1) with (2) the conductance adaptation rate is obtained (3). The right hand side of (3) provides the fundamental relationship between the conductance adaptation rate, the applied device voltage V_D and the conductance G.

$$\frac{dG}{dt} = \frac{\partial G}{\partial s} \cdot \frac{dS}{dt} \approx G \cdot \frac{2 \cdot K_p \cdot i_0}{L_0} \cdot \sinh\left(\frac{z \cdot V_D}{4 \cdot V_T}\right) \qquad (3)$$

Positive device voltages let the conductance G increase while negative voltages result in a decrease of the conductance G. Note, that a synaptic adaptation function similar to (3) has been proposed in [6], but without a specific device model at hand.

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

3 Synaptic Array

A fundamental building block which uses resistive switches as programmable synapses is illustrated in Fig. 2a. In the envisioned target architecture [1,7] integrate-and-fire neurons (IAF) are used as the neuron model.

The synaptic array receives *n* pulsed input signals X_i from a receptive field which are connected to an adjacent driver circuit *D*. This circuit transforms pulses into voltage levels for driving the resistive switches $G_{j,i}$. The wordlines are connected to the resistive switches and give rise to currents to the bitlines where the currents are summed up. At the load conductances G_L output voltages V_j are generated which are used as the input signals for IAF neurons. In order to keep the voltage swing for V_j large and power consumption low, the driver circuits operate in the so-called open-mode: only wordlines associated with active pulses ($X_i=1$) are connected to the activation voltage V_H, see Fig. 1b. As the ECM devices provide a linear I-V relationship the output voltage V_j (of column j) is given by the set of active pulses, the activated conductances and the load conductance G_L

$$V_j(t) - V_B = (V_H - V_B) \cdot (\sum_{X_i=1} G_{j,i}) / (G_L + \sum_{X_i=1} G_{j,i}) \tag{4}$$

In (4) the impact of cross current flowing through open wordlines from one bitline to the other was neglected. This is justified if at least one conductance per wordline is left in a so-called high resistive state (HRS) [7]. Due to the feedback of the load conductance G_L the transfer characteristics given by (4) is nonlinear [7]. It becomes almost linear if the probability of finding synchronous input pulses is low. This is given for uncorrelated input pulses at low pulse frequencies. Then, an input pulse X_i is exactly weighted by:

$$W_{j,i} \sim G_{j,i} / (G_L + G_{j,i}) \ . \tag{5}$$

3.1 Programming of ECM cells

Particular weights of a neuromorphic architecture are known upfront (from simulation or from calculation), such as filter coefficients of low-level feature detectors [1,7]. Here, weight programming of pre-defined values in a single time step is a sufficient operation in order to setup portions of the network. With regard to the adaptation dynamics given by (3) particular circuits can be designed which drive the conductance towards the desired value: let us consider a current source which delivers a constant but configurable current I_{prog}, cf. Fig. 2d. The current source is connected to an ECM cell with low initial conductance G_{init}. Fig. 3a shows the relationship between conductance and adaptation rate for different currents I_{prog}. If I_{prog} is positive and G is low the conductance rises in an exponential way. The voltage V_D continuously shrinks with increasing time, i.e. the adaptation rate is progressively slowed down. Finally, the adaptation rate becomes linear for a specific G dependent on I_{prog}. The turning points are marked by circles. Here, the adaptation process is considered to be virtually stopped. Finally, the specific conductance G is given by (6) using F=1 and is a *linear* function of I_{prog}. In (6) F is a design parameter which is used to identify different design options.

$$G = I_{prog} \cdot \frac{z}{4V_T} \cdot \frac{1}{F} \tag{6}$$

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

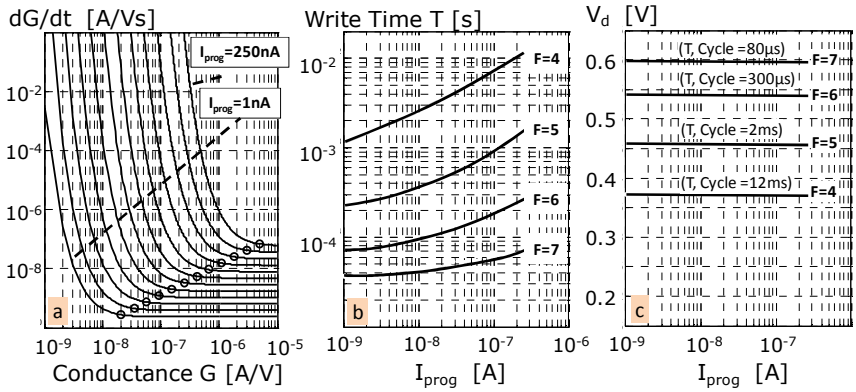

Fig. 3: a) adaptation rate; b) write time for different I_{prog}; c) V_D after write

Fig. 2c shows a simple configurable current source consisting of transistors T_1-T_3. By operating in subthreshold mode [14] the current source is able to deliver currents in the range of 1nA up to 250nA using PMOS-devices (40nm CMOS) resulting in G smaller than 10^{-6} A/V which is the appropriate range for Cu-SiO_2-based cells. In the simulations I_{prog} was specified which is mirrored 1:1 by T_3 to T_1. The drain of T_2 is connected to a wordline delivering the current to the cells which have been marked by G as well as G_P. Aim is to setup G while leaving G_P almost unchanged. In the programming scheme presented here the voltage of bitline 0 has to be set to a low level (ground for instance) while the voltage level V_{mid} on bitline 1 should be in-between V_{DD} and the lowest voltage of the wordline during write operation of G. In order to obtain a save value of V_{mid} , the current I_{prog} as well as F were varied in the simulations. First, different write times (i.e. the time T needed to obtain the desired G) are obtained, which vary between 35µs up to 12ms for a supply voltage of 0.9V, cf. Fig. 3b. For fixed F the write cycle should be large enough to ensure robust write operation for all desired input currents. Here, the maximum current I_{prog} =250nA has always the largest write time T which has been chosen as the write cycle T for the subsequent examinations. Fig.3c shows the expected final wordline voltage *after* adaptation for different values of F and currents I_{prog} (ranging from 1nA to 250nA). For larger F (i.e. shorter write cycles) the final wordline voltage becomes larger. At a nominal supply voltage of V_{DD}=0.9V the expected voltage drop across G_P can be held below 250mV (or even better for larger F) if V_{mid} is set to 650mV. Then, G_P is increased by only 0.98% (12ms write time). If the write cycle is fixed, weights which require a smaller write time than the write cycle, will become over-adapted, i.e. final conductances are larger than expected. Fig. 4a shows the effect of over-adaptation dependent on the input current for F=5 and T=2ms. For small currents I_{prog} the resulting conductances are up to 30% larger. This increase has to be considered upfront programming. Also timing errors have an effect on the adaptation. A 25% increase in the write cycle time results in an offset of approx. 5% in the conductance error, cf. Fig. 4a.

Finally, changes in the supply voltage have also an effect on the adaptation. An increase of the supply voltage V_{DD} results in an increase of the write current. Adaptation is speed up which results in over-adaptation for a fixed write cycle. The actual value of G becomes larger than the desired value G_{prog}. An increase of 11% for V_{DD} results in an offset of 3% over-adaptation, while a decrease has an effect of up to -8% over-adaptation for small currents, cf. Fig. 4b.

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

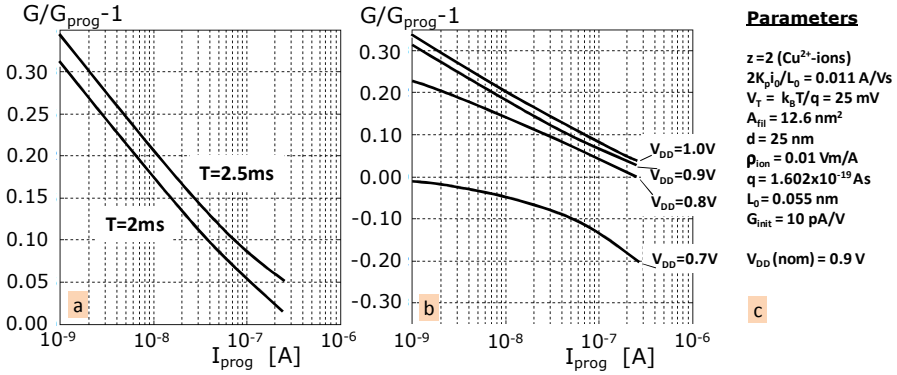

Fig. 4: Over-adaptation due to a) timing error; b) V_{DD} error; c) ECM parameters

4 Conclusion

Based on a physical device model performance figures were derived for a hybrid Cu-SiO_2-CMOS circuit with a tight I-V interaction between ECM device and active transistors. Next steps include examinations especially with respect to variability for scaled CMOS. With (3) even complex circuits can be designed which implement adaptation functions that realize weight increment/decrement with respect to particular learning laws.

References

[1] U. Ramacher, C. v.d.Malsburg, “On the Construction of Artificial Brains”, Springer, 2010.
[2] C. Bartolozzi, G. Indiveri, “Synaptic Dynamics in Analog VLSI”, Neural Computation 19, pp 2581-2603, 2007.
[3] D. Strukov, G. Snider, D. Steward, R. Williams, “The Missing Memristor Found”, Nature 53,80,2008.
[4] ITRS roadmap, http://www.itrs.net
[5] S. H. Jo, T. Chang, I. Ebong, B. B. Bhadviya, P. Mazumder and W. Lu, “Nanoscale memristor device as synapse in neuromorphic systems”, Nano Lett. pp. 1297-1301, 2010.
[6] G. Snider, “Self-organized computation with unreliable, memristive nanodevices”, Nanotechnology, vol. 18, no. 36, pp. 1-13, 2007.
[7] A. Heittmann, T.G. Noll, “Sensitivity of neuromorphic circuits using nanoelectronic resistive switches to pulse synchronization”, in Proc. ACM GLSVLSI, pp.375-378, 2011.
[8] T. Hasegawa et.al, ”Learning Abilites Achieved by a Single Solid-State Atomic Switch”, Advanced Materials, Vol.22, Iss. 16, 9.Feb. 2010.
[9] M. Versace, B. Chandler, “MoNETA: A Mind Made from Memristors”, IEEE Spectrum, Cover page fearured article, Dec. 2010.
[10] M. Aono, T. Hasegawa, “The Atomic Switch”, Proc. of the IEEE, vol.98, no.12, pp.2228-2236, Dec. 2010.
[11] S. Shin, K. Kim, and S.-M. Kang, “Compact Models for Memristors Based on Charge-Flux Constitutive Relationships,” IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 29, no. 4, pp. 590–598, 2010.
[12] S. Menzel, U. Böttger, and R. Waser, “Simulation of Multilevel Switching in Electrochemical Metallization Memory Cells”, Journal of Applied Physics Vol. 111, pp.014501-014501-5, Jan. 2012.
[13] R. Waser, “Electrochemical and Thermochemical Memories”, IEDM, pp 1-4, 2008.
[14] S.C. Liu, J. Kramer, G. Indiveri, T. Delbrück, R. Douglas, ”Analog VLSI: Circuits and Principles”, Cambridge, MA: MIT Press, 2002.