Adhesive Low Frequency Noise (LFN) in Charge Trap Transistors (CTT) for Neuromorphic Analog Processing

Alexios Birbas
Dept. of Electrical and Computer Engineering,
University of Patras,
Patras, Greece 26500
birbas@ece.upatras.gr

***Abstract*—Neuromorphic devices performing synaptic processes in the analog domain are preferable from their digital hardware counterparts for the implementation of Artificial Neural Networks (ANN). The reason is their reduced area and power requirements. Charge Trap Transistors (CTT) have been recently proposed as analog neural network computing engines due to their CMOS compatibility as well. Gate controlled differential change of the threshold voltage related to charge storage into the high-k dielectric layer (traps) is the transduction mechanism for the non-volatile device operation; the same mechanism serves the learning process when it is employed as a synaptic element. The stochastic nature of the trapping (de-trapping) procedure produces low frequency noise of RTS type and modulates the channel current. This non-desirable but unavoidable counter-intuitive effect (detrimental to the device performance) could be adhesive if this internal noise is used in order to induce stochastic resonance (SR) into the synaptic operation of the CTT. Indeed since the synaptic element is a non-linear system, SR at an engineered range of noise intensities can improve system performance (making information at the input visible at the output). This expands the learning capability of CTT into multi-layer fully connected neural networks with non linear activation functions.**

***Index Terms*— RTS noise, High-k dielectrics, neuromorhic devices, Stochastic resonance, Charge-trap transistors, Artificial Neural Networks**

I. Introduction

CTTs [1] have recently paved their way into non-volatile logic devices (memory) implementation for SoC applications. Their structure comprises an HfO based high-k metal gate (HKMG) and is fully compatible with state of the art CMOS fabrication (SOI, FinFET). This gate dielectric is known to exhibit oxygen vacancy related traps and defects which are strongly influenced by bias stress (self induced temperature) [2]. A SiO_2 interface layer is imposed between Si and the HFO_2 layer for blocking/tunneling purposes; it additionally contributes to the device good charge retention characteristics. LFN characteristics of high k-dielectrics have been studied for device quality characterization and appear to be significantly increased [3]. In this work we adopt a model for the channel current LFN calculation of a CTT in the sub-threshold region based on the generalized flat band variation [4], where a charge pumped modulated density of states $\mathrm{N_t}$ and a $\mathrm{C_{ox}}$ stemming from the equivalent two layer dielectric thickness ($\mathrm{E_{tox}}$) are incorporated (1).

$$S_{vbeeff} = \frac{q^2 \lambda k_b T N_{t,occ,trapmodulated}}{fWLC_{ox}} \tag{1}$$

Apparently the LFN noise as it appears in the channel current is direct representation of the spectral density of the flat band fluctuation and strongly dependant on the voltage bias conditions (V_{ds}, V_g). It appears that the channel current noise exhibits a 1/f behavior as expected if one considers the aggregated effect of the RTS noise originating from all the oxide traps and mobility fluctuation effects. To our interest is the noise level at the sub-threshold region assuming a charge pumping (through gate pulsing) pre-circle. Then, the oxygen traps near the conductance band are occupied (this corresponds to a gate threshold voltage increase (ΔV_{th})) while a long discharging time as well as the SiO_2 interface layer isolates those traps from the Si/SiO_2 interface.

CTTs have been proposed as analog neuromorphic computing engines [1] (convolutional neural networks for deep learning). They handle matrix multiplication (for both feed forward and error back propagation computations) by multiplying the i_{DS} subthreshold channel current with the effective $\mathrm{G_{subth}}$ off channel conductance after been modulated by the gate bias charge pulsing (trapped charge). In this work, we argue that LFN could be adhesive when CTT is employed as a neural synaptic element (where the threshold variation, due to the trap occupation, is used for learning). In general, externally induced LFN (of any color) in a neuromorphic analog synaptic element can provide an optimal level of noise (stochastic resonance of the neuron) [5] which can be beneficial in realizing a desired synaptic input-output relationship. Stochastic resonance is feasible even when internally produced noise is employed a resonance initiator[6]. This has been observed in ion- channel systems which operate alike a HMKG transistor in the sub-threshold regime.

This work has been supported by a grant (10073 FPGAs4SGs) of Western Region of Greece to the University of Patras under the RIS-3 program.

II. ARTIFICIAL NEURAL NETWORKS

Neural networks realize non-linear functions to provide optimal input-output relations as shown in Fig.1. In particular cases neural nets comprise hidden units employing threshold elements which in turn can be affected by noise. These threshold elements allow the implementation of threshold activation functions and in generally can be implemented by digital hardware. Activation functions can be replaced during learning by sigmoids and used as threshold functions at neural net runtime. In general, smoothing out the activation functions is a necessity during the learning period for performing gradient descent algorithms. It has been investigated that Stochastic Resonance (SR) can be very beneficial when the input-output relation is nonlinear in the sense that adding noise helps smoothing out the activation function. Of special interest in this work is the examination of threshold elements which's inputs are modified by random fluctuations (noise) during ANN training and run time. The noise is not just added to the input-output pairs which train the network but instead are used to compute a distribution from which new noisy data are generated. In the implementation of a large NN on silicon, binary outputs with steep threshold functions are convenient in terms of the ease of construction.

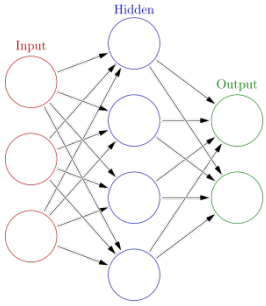

Fig. 1. Neural Network with hidden layers

The NNs in general face difficulties to convey with software implementations (given that the computational power of a human brain ranges at 10^{11} MFLOPs for a limited power budget of 10 W) so hardware implementations are heavily pushed alternatively. However, a new class of neuromorphic circuits have been lately evolved which can closer emulate the neurons operability and exploit their analog nature. The ANN general structure comprises a synapse (network wiring, synaptic weights and learning) and the neuron (neuron state-summation of weighted inputs and the activation function which is highly non-linear).

III. CTT

Charge trap transistors (CTTs) have been recently proposed as a compact continuously tunable non-volatile synapse devices (CMOS-compatible analog neural network computing engine) [7].

It is feasible today to fabricate CTTs in 14nm FinFET technology without extra process complexity and masks as

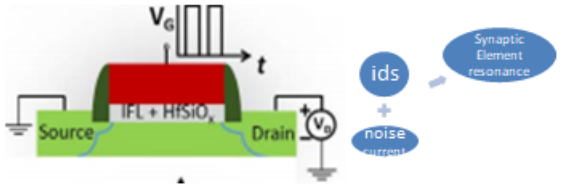

Fig. 2. A CTT structure. A gated pulse is enabling the synaptic operation of the transistor. Convolution of channel current and noise current can lead to stochastic resonance

shown in Fig.2. The enhanced and stabilized charge-trapping behavior of CTTs, allows their use into analog non-volatile memory (trap charge dissipates very slowly) and make them candidates for analog computing (neuromorphic) neural networks (learning). Device key characteristic is an interfacial layer SiO_2 followed by an $HfSiO_x$ layer as the high-k gate dielectric, common in modern CMOS technologies. The device gate threshold voltage V_{th} is modulated by the trapped charge (long trapping and de-trapping procedure) into the gate dielectric depending on a pulse train applied on the gate. Such a device exhibits very low energy consumption per synaptic operation.

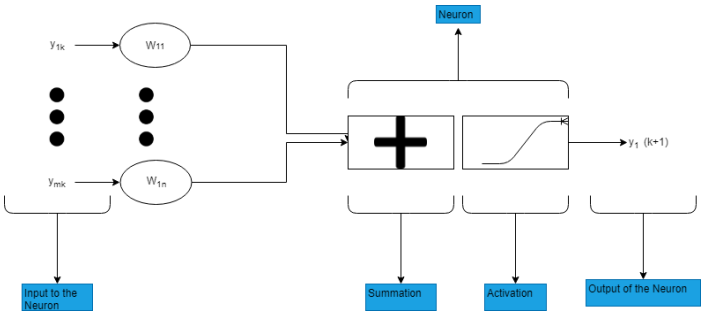

Fig. 3. A generic synapse part of a fully connected multiplication array where input data are fed (as voltages) at the drain of the CTTs while the weights are fed to the gates of the CTTs as pulse train gate voltages

As shown in Fig.3, the CTT matrix implementing a synapse connects the output $Y_j(k+1)(j = 1 \dots N)$ with the input $Y_i(k)(i = 1 \dots M)$ through a weighted MxN matrix with synaptic weight coefficients W_{ij}:

$$Y_j(k+1) = \alpha \sum_{t}^{M} Y_i(k) W_{ij} \tag{2}$$

The CTTs are biased in the sub-threshold regime (V_{GS}-V_{DS} $< V_{th}$). The fast-reading and the slow nature of CTT discharging allows programming by storing the synaptic weights in the threshold voltages (through positive/trapping and negative/de-trapping voltage pulse trains.) A resistor connected to the sources of all the transistors in a row allows the collection of all the drain to source currents thus implementing the summation producing the $Y_j(k+1)$ output. If the data

are available at the particular nodes at the same time the calculation is performed in one circle given that $V_{th,ij}$ has been pre-trained with the appropriate pulse train. For a given node, the inputs are multiplied by the weights in a node and summed together. This node value is referred to as the summed activation of the node. The summed activation is then transformed via an activation function (α) and defines the specific output or activation of the node. The simplest activation function is referred to as the linear activation, where no transform is applied at all. A network comprised of only linear activation functions is very easy to train, but cannot learn complex mapping functions. Linear activation functions are still used in the output layer for networks that predict a quantity (e.g. regression problems). Nonlinear activation functions are preferred as they allow the nodes to learn more complex structures in the data. Ikemoto et.al [5] have recently propose non linear activation functions (threshold elements) that are modified by noise. The exploitation of the noise in this work is fundamentally different than the addition of noise to the input - output pairs for training purposes. This is a demonstration of noise assisted switching, wherein both the noise and the non-linear characteristics are exploited in order to enhance the flow of information to the output. RTN noise induced by single oxide trap is observed at the drain current of FETs with high-k dielectrics [8]. Normally the single-trap induced RTN is a typical Poisson process and the stochastic trapping/detrapping can be characterized by the capture time constants τ_c and τ_e.

IV. STOHASTIC RESONANCE

Stochastic resonance (SR) is a mathematical concept that in general refers to the increased sensor sensitivity of a system due to the external addition of a finite level of noise [4][5][10][11]. Generally a system under SR exhibits a maximum SNR upon an optimal noise level applied. A system with an internal noise level can exhibit SR upon inducing an external noise source into the system; this external noise source eventually induces SR. The system that exhibits SR will attain a maximum when an optimal level of noise is applied for a specific frequency band. Detecting signals buried in noise in FETs with strong current nonlinearity which overcomes the thermal limitation and dynamic bistability is achieved by adding common noise into the FET operating at the subthreshold region under a pulsed voltage [9]. The FET is driven by an external pulse train and a noise source. However, it has been shown that if the physical mechanisms producing internal noise are well defined and understood it would be possible to engineer the system so that the internal noise alone is able to attain SR. The effect has been observed in nanoscale biosensors (ion channel switch biosensors) which uses ions to transduce events of molecular recognition into detectable changes in impedance [6]. To improve the sensitivity of these sensors, one aims to lower the sensors limit of detection. This is accomplished by enhancing the SNR. It is worth mentioning that the internal noise in these systems arises from the probabilistic gating of the ion channels themselves. In modeling these stochastic fluctuations as a noisy current component in the Hodgkin-Huxley equations, it has been shown that it first increases with noise, attains a maximum, and then asymptotically approaches zero, thus showing the signature of SR [6]. It is evident that one can achieve SR in such devices if the gated transductive device is properly engineered. In this work we employ CTT as a neuromorphic device for implementing ANNs. CTTs are CMOS compatible nanowire transistors driven by gate voltage train pulses aiming at trapping and detrapping of transistors into the dielectric. This gated structure exhibits SR [9] under the supply of external noise applied on the gate of the transistor along with a voltage train pulse with a specific offset. A correlation coefficient C is defined as a metric to evaluate the information transmitted from input to output under stochastic resonance. By defining the C coefficient it has been shown that this coefficient increases and then decreases as a function of noise an indication that SR is achieved for a certain input noise as shown in Fig.4. The stochastic resonance is attained to the bistable energy system diagram providing hysteresis characteristics and represents fluctuations of the barrier height.

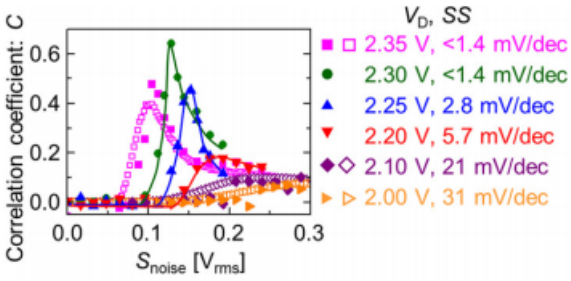

Fig. 4. The correlation coefficient C as a function of the increasing induced noise indicating stochastic resonance at different voltages and sub-threshold slopes (The Fig. has been reproduced from K. Nishigusi and A. Fujiwara ref. [9]

V. DISCUSSIONS AND CONCLUSIONS

As was shown in section III, the inference and training operations rely on the matrix multiplication for either feed forward or error back propagation. Indeed the specific CTT operating at the sub-threshold regime executes the multiplication:

$$V_{oj} = R_{source} K_n \frac{W}{L} \sum_{t}^{M} V_{ds,ij}.(V_{gs} - V_{th,ij}) \qquad (3)$$

$V_{ds,ij}.(V_{gs}\text{-}V_{th,ij})$ can be relaxed to a direct product $V_{ds,ij}$,$V_{th,ij}$ when an input related offset is removed at the digital domain post processing [7]. R_{source} is the source series resistance connected to the CTTs sources of the entire row. V_{oj} then corresponds to the $Y_j(k+1)$ neural cell and it is basically calculated as the $R_{source}.G_{subth}$ product where G_{subth} is the off subthreshold conductance. $V_{th,ij}$ is the threshold voltage which incorporates the threshold variation ΔV_{th} associated with the trapping density change during

initialization (trapping), programming (trapping) and erase (de-trapping) phases of the neural net training. ΔV_{th} depends both on the self heating (applied V_{ds}) but mostly on the V_{gs} pulse train (voltage level and pulse frequency). It also depends on the trap location (l-distance from the channel). Moreover, it is well understood that LFN in nanoscale transistors, in the linear regime of operation, is associated with the stochastic oxide trap fluctuating occupancies. This holds true even for planar high/k metal gate MOSFETS with ultra thin HfO_2 gate dielectrics (i.e CTTs). This low frequency noise is of the RTS type (RTN) and a single-trap induced RTN comprises both a capture time constant τ_c and an emission time constant τ_e representing the average waiting time (to capture or emit) a carrier at the trap level from (to) the channel. RTN induced by N traps give a colored noise spectral density for the channel current and the noise dependence on V_g is determined by:

$$\frac{l}{E_{ox}} = \frac{K_b T}{q} \frac{\partial ln(\frac{\tau_c}{\tau_e})}{\partial V_g} \tag{4}$$

$$E_{ox} = E_{si} + \frac{\epsilon_{si}}{\epsilon_{hf}} E_{hf} \tag{5}$$

E_{si} and E_{hf} are the SiO_2 and HfO_x thicknesses respectively. Equations (1), (4) and (5) indicate the close relation of RTN with the gate voltage pulse train (through the modulation of the effective trap density) and with the dielectric characteristics. Similarly, V_{th} associated with the threshold voltage variation (during gate charging and erasing periods neural net learning), depends on the same parameters. The RTN noise peak current density in such devices can be as high as 6-7 % of the channel current and can even exhibit anomalous RTN behavior where RTN data exhibit two zones with identical amplitudes but reversal time constants (spectral density) for the source current and the noise [8]. Recalling back from section IV the discussion regarding stochastic resonance it is evident that in a CTT, the RTN is the internal dominant noise mechanism and under certain device conditions could induce stochastic resonance in the CTT. As was explained, both phenomena depend on the same parameters which mean that it is possible to engineer the CTT so that the input signal of the non-linear system (small to affect the output) becomes observable by adding appropriately the internal non-zero level noise to the system. Engineering parameters such as E_{ox}, capturing constants, gate pulse characteristics (amplitude and frequency) and gate offset level can eventually create a CTT operating in a learning process under stochastic resonance conditions. The quantitative analysis needed is currently under research. In [11] it has been shown that regardless the fact that the external noise source, usually employed in order to induce stochastic resonance in a neural network, is white noise, the employment of a colored noise source (pink -1/f noise) can amplify the output signal by orders of magnitude. The synaptic CTT can then used for performing the synaptic calculations and the same time SR can provide the threshold activation functioning required by the hidden layers of the NN. This could also be achieved by using a separate CTT as a pass-transistor driven by the same gate voltage pulse train substituting the source resistance of the synaptic CTT. In conclusion: this work provides evidence that a Charge Trap Transistor is a suitable neuromorphic device for the expedition of synaptic calculations in the analog domain. The internally produced RTS noise associated with trapping and de-trapping of charges into the SiO_2/HfO_2 dielectric system is closely associated with the learning procedure since both are related to charge trapping in the high-k dielectric traps. The internally produced noise can eventually induce SR into the CTT system operating as a synapse calculation engine of the neural net. This SR, in turn, improves (decreases) the information loss of the non-linear system (CTT) by making it closer to a linear system.

REFERENCES

[1] F. Khan, E. Cartier, J. C. S. Woo and S. S. Iyer, "Charge Trap Transistor (CTT): An Embedded Fully Logic-Compatible Multiple-Time Programmable Non-Volatile Memory Element for High- k -Metal-Gate CMOS Technologies," in IEEE Electron Device Letters, vol. 38, no. 1, pp. 44-47, Jan. 2017.

[2] Kothandaraman , X. Chen , D. Moy , D. Lea , et. al , Proc. 2015 IEEE International Reliability Physics Symposium

[3] Maryam Olyaei, Doctoral Thesis Technology KTH Royal Institute of Technology Stockholm, Sweden 2015

[4] E. G. Ioannidis, C. G. Theodorou, T. A. Karatsori, S. Haendler, C. A. Dimitriadis, and G. Ghibaudo, Drain-current flicker noise modeling in nMOSFETs from a 14-nm FDSOI technology, IEEE Transactions on Electron Devices, vol. 62, no. 5, pp. 15741579, 2015.

[5] Shuhei Ikemoto, Fabio DallaLibera, Koh Hosoda, Noise-modulated neural networks as an application of stochastic resonance, Neurocomputing, Volume 277, 2018, Pages 29-37,

[6] E. Stava, S. Choi, H-S. Kim, and R. H. Blick, On-Chip Stochastic Resonance of Ion Channel Systems With Variable Internal Noise, IEEE Transactions On Nanobioscience, Vol. 11, No. 2 (2012), pp. 169-175

[7] Y. Du et al., "An Analog Neural Network Computing Engine using CMOS-Compatible Charge-Trap-Transistor (CTT)," in IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems.

[8]] S. Guo, R Wang, D. Mao, Y. Wang and R. Huang, Anomalous random telegraph noise in nanoscale transistors as direct evidence of two metastable states of oxide traps, Scientific Reports, Vol. 7, No: 6239 (2017)

[9] K. Nishiguchi and A. Fujiwara Detecting signals buried in noise via nanowire transistors using stochastic resonance Appl. Phys. Lett. 101, 193108 (2012)

[10] P. Krauss, C. Metzner, A. Schilling, C. Shultz, K. Tziridis, B. Fabry and H. Schule , Adaptive stochastic resonance for unknown and variable input signals, Scientific Reports Vol. 7, No 2450 (2017)

[11] A. Castellanos, Stochastic Resonance in neural network, noise color effects, arXiv:1810.06731 [nlin.AO]