Title
Unsupervised Learning Using Charge-Trap Transistors

Journal
IEEE Electron Device Letters, 38(9)

Authors
Gu, Xuefeng
Iyer, Subramanian S.

Publication Date
2017-07-01

Peer reviewed

Unsupervised Learning Using Charge-Trap Transistors

Xuefeng Gu and Subramanian S. Iyer, *Fellow, IEEE*

***Abstract*— Unsupervised learning is demonstrated using a device ubiquitously found in today's technology: a transistor with high-k-metal gate. Specifically, the charge-trapping phenomenon in the high-k gate dielectric is leveraged so that the device can be used as a non-volatile analog memory. Experimental data from 22-nm SOI devices reveal that a charge-trap transistor possesses promising characteristics for implementing synapses in neural networks, such as very fine tunability, weight-dependent plasticity, and low power consumption. A proof-of-concept winner-takes-all neural network is simulated based on experimental data and perfect clustering is achieved within tens of training cycles. This means that the network can be trained for multiple times, and a larger system can be built. The robustness of the procedure to the device variation is also discussed.**

***Index Terms*—High-k-metal gate, charge-trapping, unsupervised learning, neuromorphic computing**

I. Introduction

A compact and continuously tunable non-volatile synapse device is essential for biologically inspired intelligent systems, which promise to be much more power- and time-efficient than conventional von-Neumann architectures [1−6]. Over the years there has been an expanding group of candidates proposed for analog synapses, among which are resistive memory (ReRAM) and phase-change memory (PCM) [7−11]. These emerging memory devices have been used in neural networks for both supervised and unsupervised learning [6, 9−11]. Besides the complexities of introducing new materials and processes, their statistical operating mechanisms lead to challenging variation issues. Device endurance is an additional concern. For example, a typical ReRAM shows a conductance spanning more than two orders of magnitude within first 100 programming cycles at identical programming conditions [7]. Devices based on charge-trapping include floating-gate transistors [12], transistors with an organic gate dielectric [13], and carbon nanotube transistors [14]. However, none of these proposals are both fully CMOS-compatible (in terms of process and operating voltage) and manufacturing-ready.

The charge-trapping phenomenon in a transistor (hence charge-trap transistors, CTT) with high-k-metal gate has traditionally been considered a reliability concern, causing bias

This work was supported in part by DARPA (FA8650-16-1-7648).

X. Gu and S. S. Iyer are with the Center for Heterogeneous Integration and Performance Scaling (CHIPS), Electrical Engineering Department, University of California, Los Angeles, Los Angeles, CA, 90095, USA (xfgu@ucla.edu).

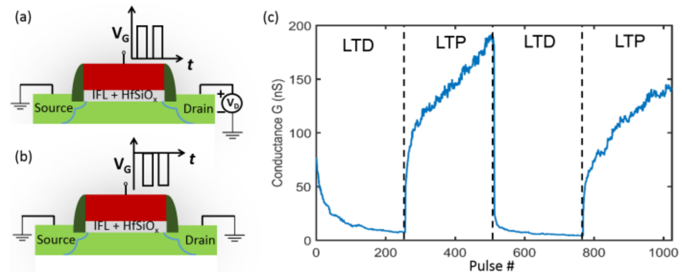

Fig. 1. Configurations of the CTT in the (a) LTD and (b) LTP regimes. (c) Reversible and reproducible device conductance change through four cycles.

temperature instability, etc. However, it was recently discovered that, with a drain bias during the charge-trapping process, many more carriers can be trapped in the gate dielectric very stably, and more than 90% of the trapped charge can be retained after 10 years even when the device is baked at 85 ℃ [15]. This enhanced and stabilized charge-trapping behavior has been discussed in detail in [15] and successfully exploited for embedded non-volatile digital memory applications [16, 17].

The CTT may also be used to realize a non-volatile analog memory. In this Letter, we demonstrate how a transistor with high-k gate dielectric, specifically $HfSiO_x$, can be configured as an analog synapse. These synapses can be used in neural networks to implement both supervised and unsupervised learning. Here, as an example, we demonstrate the implementation of unsupervised learning in a neural network using CTT as the plastic synapses. We first investigate the characteristics of the CTT that are essential to the implementation of unsupervised learning in neural networks. Very fine tunability and weight-dependent plasticity are experimentally demonstrated using commercial 22-nm SOI devices. A low power consumption of ~ nJ per synaptic operation is also estimated. An unsupervised-learning winner-takes-all (WTA) neural network featuring CTTs as the plastic synapses is then simulated based on experimental data. Results show that the system learns rapidly in a few tens of training cycles, which allows for multiple learning cycles well within the endurance limits of the CTT. Furthermore, we show that the WTA algorithm taking advantage of the inherent properties of CTTs is robust to device variation.

II. Experimental Details and CTT Characteristics

N-type CTTs with an interfacial layer (IFL) SiO_2 followed by an $HfSiO_x$ layer as the gate dielectric are used in this study. It should be noted that, although this demonstration features planar SOI devices, the mechanisms apply to bulk

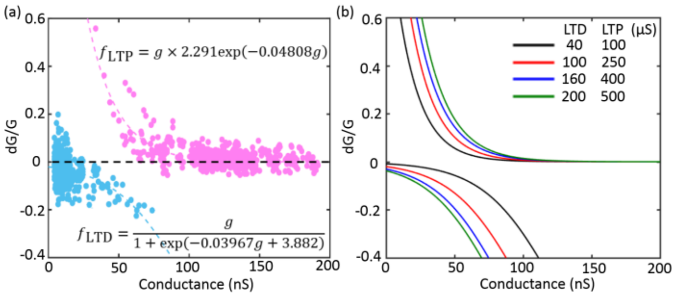

Fig. 2. (a) The weight-dependent plasticity when five trapping/detrapping pulses are applied in the LTD/LTP regimes, respectively. (b) Fitted curves when pulses of different widths are applied.

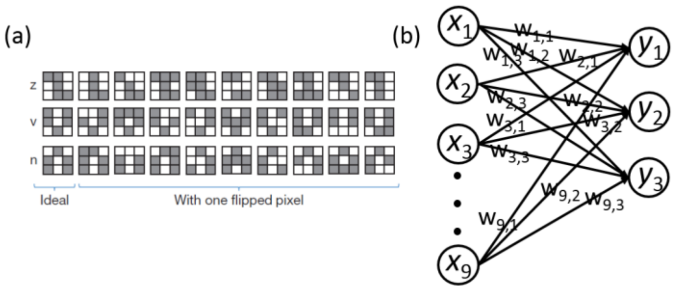

Fig. 3. (a) Stylized letters z, v, n, and one-bit-flipped noisy versions of them (adapted from Ref. [18]). (b) The setup of the unsupervised neural network.

substrates/FinFETs as well. The subthreshold OFF-conductance (G_{OFF}) of the CTT at V_{DS} = 50 mV and V_{GS} = 0 will be used as the synaptic weight throughout this Letter. In the operation of a CTT-based synapse, its G_{OFF} is modified by changing the amount of charge trapped in the high-k layer and thus shifting the threshold voltage (V_T) of the transistor. In the long-term depression (LTD) regime, a positive gate pulse is applied and electrons are trapped into $HfSiO_x$ through the IFL, increasing V_T and decreasing G_{OFF} (Fig. 1(a)); in the long-term potentiation (LTP) regime, a negative gate pulse is applied and trapped electrons tunnel back into the channel, decreasing V_T and increasing G_{OFF} (Fig. 1(b)). In our experiments, a CTT is first pre-programmed to an intermediate starting state by applying a gate pulse of 2.5 V for 60 µs with V_D = 1.3 V. The device subsequently goes through four cycles: two LTD and two LTP cycles, with 256 trapping or detrapping pulses in each cycle. In the LTD cycle, G_{OFF} is decreased by a 20-µs, 2.5 V gate pulse with V_D = 1.3 V; in the LTP cycle, G_{OFF} is increased by a 50-µs, −2.6 V gate pulse with zero drain bias. The resulting G_{OFF} is shown in Fig. 1(c) where a reversible and reproducible modification of synaptic weights can be observed. Over 200 levels are achieved for both LTP and LTD regimes with a fine resolution of less than 1 nS for LTP and 0.25 nS for LTD. As we will show later, although the LTD has a smaller dynamic range, it will not affect the convergence of the learning algorithm.

An important characteristic of CTTs when used as analog synapses is the weight-dependent plasticity: at different G_{OFF}, the effect of programming pulses on G_{OFF} is different. The weight-dependent plasticity is also found in biological synapses, and might be interesting to emulate the brain. Shown in Fig. 2(a) is the relative G_{OFF} change as a function of G_{OFF} itself when five trapping and detrapping pulses as specified above are applied. It is observed that, in the LTP regime, the relative G_{OFF} increase is smaller when the initial G_{OFF} is larger; on the contrary, in the

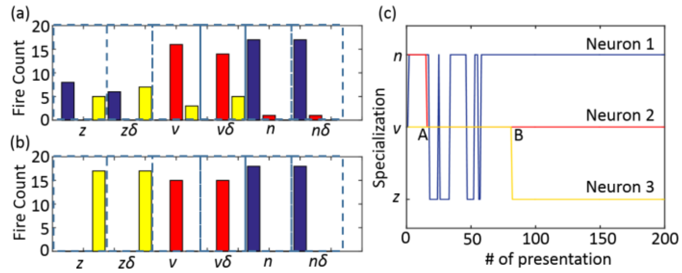

Fig. 4. Fire counts from three output neurons (a) before and (b) after training. Blue, red, yellow: output neurons 1, 2, and 3. "δ" denotes a noisy version. (c) The evolution of the output neuron specializations as the network is trained.

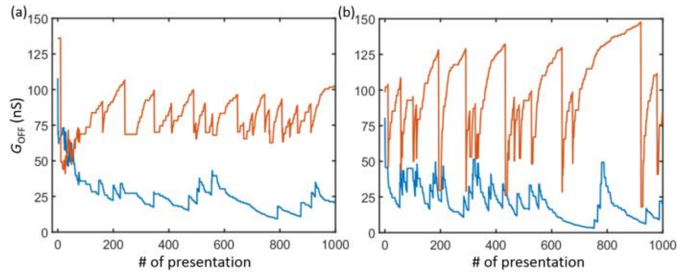

Fig. 5. An example of the evolution of synaptic weights $G_{\mathrm{OFF}1,1}$ (blue) and $G_{\mathrm{OFF}2,1}$ (red) for different programming times: (a) Two pulses are applied for LTD/LTP, and (b) Five pulses are applied for LTD/LTP.

LTD regime, the relative G_{OFF} reduction is larger when the initial G_{OFF} is larger. The curves corresponding to the LTP and LTD regimes are fitted to exponential and sigmoid functions, respectively, for different programming times (Fig. 2(b)). As expected, a longer programming time consistently leads to a larger G_{OFF} change because of the larger V_T change caused by more trapped/detrapped charge [16].

The energy consumption in the LTP regime is minimal since it is only due to electrons being detrapped from the high-k layer. In the LTD regime, the energy dissipation is mainly through the channel current because of the drain bias; it is given by $E = V_{\mathrm{DS}} \int I_{\mathrm{D}} \cdot dt$ where I_D is the channel current. For a device with a W/L = 20 nm / 20 nm and programming conditions given above, E is estimated to be 0.5 nJ. This is a reasonable value compared to the range of pJ to hundreds of nJ reported for many other synapse candidates [10].

III. Theory and Simulation

CTTs are next used as synapse devices in a one-layer WTA neural network aiming at classifying stylized letters z, v, n, and one-bit-flipped noisy versions of them (Fig. 3(a)) [18]. The input layer of the network has nine neurons corresponding to nine pixels of the pattern and the output layer has three neurons corresponding to the three categories: z, v, and n, respectively (Fig. 3(b)). For each output neuron j (1, 2, or 3), its output is determined by $y_j = \sum_{i=1}^{9} x_i G_{\mathrm{OFF}i,j}$, where $G_{\mathrm{OFF}i,j}$ is the G_{OFF} of the CTT between the input neuron i and the output neuron j, and x_i is the input which is 50 mV when the ith pixel is black (firing) or 0 when the ith pixel is white (not firing). For each presentation of a pattern, the neuron with the largest output fires and claims the pattern, and only the 9 synaptic weights associated with this neuron are updated with a WTA rule [19].

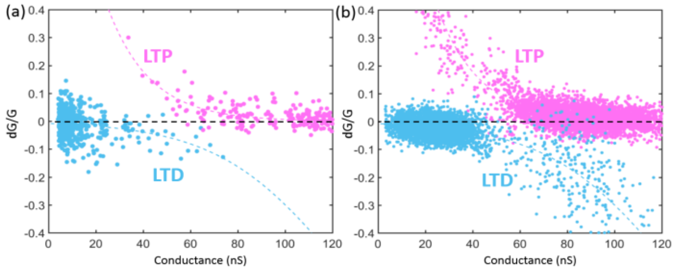

Fig. 6. (a) Experimentally measured and (b) Empirically determined relative conductance change as a function of the conductance itself in the LTP and LTD regimes. The algorithm converges with the variation shown in Fig. (b).

Specifically, only when output neuron j has the largest output and fires (wins) are $G_{OFFi,j}$ (i = 1−9) updated: $\Delta G_{OFFi,j}$ is increased by detrapping pulses if the input neuron i also fires or decreased by trapping pulses if the input neuron i does not fire. In the simulation, we start from CTTs with random G_{OFF} ranging from 50−150 nS. Training of the neural network starts with randomly selecting a pattern from z, v, or n, and presenting it to the network. Then a random bit of the pattern is flipped and the noisy version is presented to the network again. Formulas fitted from experimental data is used to update the synaptic weights. The entire process is free of any intervention.

IV. Results and Discussion

In the simulation, a total of 1000 patterns are presented to the neural network with 500 correct ones and 500 noisy ones. Two trapping and detrapping pulses as specified above are applied during the LTD and LTP regimes. Figs. 4(a) and (b) show the clustering results for the first and the last 100 presentations, respectively. It is observed that a substantial number of misclassifications occur in the first 100 cycles, while all patterns are correctly classified for the last 100 cycles. To better understand the convergence behavior of the algorithm, a specialization function, S_i, is defined for each output neuron i, as the pattern **x** (z, v, or n) which yields the largest output y_i for the neuron. Perfect clustering is achieved when the neuron specializations remain constant and correspond to three different patterns as the neural network is trained. Fig. 4(c) shows the specializations of the output neurons as the network is trained. In fact, perfect clustering is achieved after only 82 training cycles, after which Neurons 1, 2, and 3 correspond to patterns n, v, z, respectively. Between points A and B, even though the specializations of Neurons 2 and 3 stay constant, the algorithm is not convergent since both neurons claim the letter v. It should be further noted that this example is only to illustrate the evolution of specializations and does not represent a typical case. It is verified through 10,000 simulation runs that, the average number of cycles after which perfect clustering is achieved is only 24, well within the demonstrated endurance of over 1,000 for CTT-based non-volatile memory [16].

Fig. 5 depicts an example of the evolution of the synaptic weights $G_{OFF1,1}$ and $G_{OFF2,1}$. It is observed that, the sharp decreases in $G_{OFF2,1}$ are larger than the sharp increases in $G_{OFF1,1}$, which is caused by the asymmetry between LTP and LTD found in Fig. 1(c). It is also observed that, the weights, starting from random values, eventually reach a steady state after which each weight only varies around a certain value. In this example, the steady-state is 23.8 nS for $G_{OFF1,1}$ and 93.2 nS for $G_{OFF2,1}$ for the last 100 cycles when two trapping/detrapping pulses are applied in the LTD/LTP regimes. These two values, representing respectively "low" and "high" weights after training, vary with the applied programming conditions. For instance, when five trapping/detrapping pulses are applied, a "low" of 15.2 nS and a "high" of 95.8 nS are obtained. When a longer programming pulse is applied, larger G_{OFF} change is induced in each update step, leading to higher "high" and lower "low" eventual weights. Larger weight changes also result in faster convergence and a smaller noise margin. It is anticipated that the amplitudes of the trapping/detrapping pulses will have similar effects on the convergence behavior.

In practice, when actual CTTs are used to construct the neural networks, the effect of device variation on the robustness of the algorithm needs to be evaluated. We illustrate here the example where two trapping and detrapping pulses are used to update the weights (Fig. 6(a)). An empirically determined variation of Gaussian distribution with 3σ of $f_{10pulse} - f_{2pulse}$ is added to the conductance change calculated from fitted equations, where $f_{10pulse}$ denotes the fitted conductance change when ten pulses are used to update the weights and f_{2pulse} denotes the fitted conductance change when two pulses are used to update the weights. More variation is introduced when $G_{OFF} > 60$ nS in the LTP regime and when $G_{OFF} < 40$ nS in the LTD regime to better approximate the experimental data. With this variation, the simulation was performed for 10,000 times and a 100% perfect clustering rate was achieved. Fig. 6 (b) depicts an example of ΔG_{OFF} as a function of G_{OFF} itself from one of these simulation runs. It is indeed observed that the conductance change with the empirically introduced variation is comparable to the experimental data. With this methodology, it is also found that a longer programming time leads to a less robust algorithm: perfect clustering cannot be achieved when five LTP/LTD pulses are applied. It means that the effects of the variation are smaller when the programming time is shorter. This is because a shorter programming time corresponds to a smaller ΔG_{OFF} in each update step.

V. Conclusion

We have shown that, the CTT, as a nonvolatile analog memory device, exhibits intriguing properties for brain-inspired computing. A proof-of-concept WTA neural network featuring CTTs as its synapses is presented to cluster stylized letters. The number of training cycles required to achieve perfect clustering is well within demonstrated endurance of CTT. The convergence behavior of the algorithm varies with different programming conditions, and the algorithm is robust to device variation. These findings pave the way to an ultra-large scale, completely CMOS-based intelligent system without any material or process complexities.

References

We thank Faraz Khan from UCLA for helpful discussions. We also thank UCOP for their support (MRP-17-454999).

References

[1] D. S. Modha, R. Ananthanarayanan, S. K. Esser, A. Ndirango, A. J. Sherbondy, and R. Singh, "Cognitive Computing," *Communications of the ACM*, vol. 54, pp. 62-71, Aug. 2011. doi: 10.1145/1978542.1978559

[2] S. K. Esser, P. A. Merolla, J. V. Arthur, A. S. Cassidy, R. Appuswamy, A. Andreopoulos, D. J. Berg, J. L. McKinstry, T. Melano, D. R. Barch, C. di Nolfo, P. Datta, A. Amir, B. Taba, M. D. Flickner, and D. S. Modha, "Convolutional Networks for Fast, Energy-efficient Neuromorphic Computing," *Proc. NAS*, vol. 113, 11441-11446, Aug. 2016. doi: 10.1073/pnas.1604850113

[3] S. B. Furber, "Brain-inspired Computing," *IET Computers & Digital Techniques*, vol. 10, pp. 299-305, Jan. 2016. doi: 10.1049/iet-cdt.2015.0171

[4] A. Calimera, E. Macii, and M. Poncino, "The Human Brain Project and neuromorphic computing," *Funct. Neurol.*, vol. 28, pp. 191-196, Oct. 2013. doi: 10.11138/FNeur/2013.28.3.191

[5] P. A. Merolla, J. A. 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, and D. S. Modha, "A million spiking-neuron integrated circuit with a scalable communication network and interface," *Science*, vol. 345, pp. 668-673, Aug. 2014. doi: 10.1126/science.1254642

[6] P. Yao, H. Wu, B. Gao, S. B. Eryilmaz, X. Huang, W. Zhang, Q. Zhang, N. Deng, L. Shi, H.-S. P. Wong, and H. Qian, "Face classification using electronic synapses," Nature Communications, vol. 8, pp. 15199.1-15199.8, May 2017. doi: 10.1038/ncomms15199

[7] S. Yu, Y. Wu, R. Jeyasingh, D. Kuzum and H. S. P. Wong, "An Electronic Synapse Device Based on Metal Oxide Resistive Switching Memory for Neuromorphic Computation," *IEEE Transactions on Electron Devices*, vol. 58, no. 8, pp. 2729-2737, Aug. 2011. doi: 10.1109/TED.2011.2147791

[8] S. Mandal, A. El-Amin, K. Alexander, B. Rajendran, and R. Jha, "Novel synaptic memory device for neuromorphic computing," *Scientific Reports*, vol. 4, pp. 5333.1-5333.10, Jun. 2014. doi: 10.1038/srep05333

[9] S. Kim, M. Ishii, S. Lewis, T. Perri, M. BrightSky, W. Kim, R. Jordan, G. W. Burr, N. Sosa, A. Ray, J.-P. Han, C. Miller, K. Hosokawa, and C. Lam, "NVM neuromorphic core with 64k-cell (256-by-256) phase change memory synaptic array with on-chip neuron circuits for continuous in-situ learning," *2015 IEEE International Electron Devices Meeting (IEDM)*, Washington, DC, pp. 17.1.1-17.1.4, Dec. 2015. doi: 10.1109/IEDM.2015.7409716

[10] D. Kuzum, S. Yu, and H. S. P. Wong, "Synaptic Electronics: Materials, Devices, and Applications," *Nanotechnology*, vol. 24, 382001.1-22, Sept. 2013. doi: 10.1088/0957-4484/24/38/382001

[11] S. B. Eryilmaz, E. Neftci, S. Joshi, S. B. Kim, M. BrightSky, H.-L. Lung, C. Lam, G. Cauwenberghs, and H.-S. P. Wong, "Training a Probabilistic Graphical Model with Resistive Switching Electronic Synapses," IEEE Transactions on Electron Devices, vol. 63, no. 12, pp. 5004-5011, Dec. 2016. doi: 10.1109/TED.2016.2616483

[12] D. Hsu, M. Figueroa, and C. Diorio, "Competitive learning with floating-gate circuits," *IEEE Transactions on Neural Networks*, vol. 13, pp. 732-744, May 2002. doi: 10.1109/TNN.2002.1000139

[13] M. Debucquocy, M. Rockelé, J. Genoe, G. H. Gelinck, and P. Heremans, "Charge trapping in organic transistor memories: On the role of electrons and holes," Organic Electronics, vol. 10, pp. 1252-1258, Jul. 2009. doi: 10.1016/j.orgel.2009.07.005

[14] S. Kim, J. Yoon, H.-D. Kim, and S.-J. Choi, "Carbon Nanotube Synaptic Transistor Network for Pattern Recognition," *ACS Appl. Materials & Interfaces*, vol. 7, pp. 2549-25486, Oct. 2015. doi: 10.1021/acsami.5b08541

[15] F. Khan, E. Cartier, C. Kothandaraman, J. C. Scott, J. Woo, and S. S. Iyer, "The impact of self-heating on charge trapping in highk -metal-gate nFETs," *IEEE Electron Device Lett.*, vol. 37, no. 1, pp. 88–91, Jan. 2016. doi: 10.1109/LED.2015.2504952

[16] 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," *IEEE Electron Device Letters*, vol. 38, no. 1, pp. 44-47, Jan. 2017. doi: 10.1109/LED.2016.2633490

[17] J. Viraraghavan, D. Leu, B. Jayaraman, A. Cestero, R. Kilker, M. Yin, J. Golz, R. R. Tummuru, R. Raghavan, D. Moy, T. Kempanna, F. Khan, T. Kirihata, S. S. Iyer, "80Kb 10ns read cycle logic Embedded High-K charge trap Multi-Time-Programmable Memory scalable to 14nm FIN with no added process complexity," 2016 IEEE Symposium on VLSI Circuits (VLSI-Circuits), Honolulu, HI, pp. 1-2, Jun. 2016. doi: 10.1109/VLSIC.2016.7573462

[18] M. Prezioso, F. Merrikh-Bayat, B. D. Hoskins, G. C. Adam, K. K. Likharev, and D. B. Strukov, "Training and Operation of An Integrated Neuromorphic Network Based on Metal-oxide Memristors," *Nature*, vol. 521, pp. 61-64, May 2015. doi: 10.1038/nature14441

[19] A. Serb, J. Bill, A. Khiat, R. Berdan, R. Legenstein, and T. Prodromakis, "Unsupervised learning in probabilistic neural networks with multi-state metal-oxide memristive synapses," vol. 7, pp. 12611.1-12611.9, Sept. 2016. doi: 10.1038/ncomms12611