A Novel Architecture to Build Ideal-linearity Neuromorphic Synapses on a Pure Logic FinFET Platform Featuring 2.5ns PGM-time and 10^{12} Endurance

E. R. Hsieh[1,3], H. Y. Chang[1], Steve S. Chung[1], T. P. Chen[2], S. A. Huang[2], T. J. Chen[2], Osbert Cheng[2], and S. Simon Wong[3]
[1]*Dept. of Electronics Eng. & Institute of Electronics, National Chiao Tung University, Taiwan,* [2]*United Microelectronics Corporation (UMC), Taiwan,*
[3]*Department of Electrical Engineering, Stanford University, Stanford, CA 94305, USA*

Abstract- In this work, we will explore pure logic FinFET devices to realize the functionality of linear weight tuning capability as electric synapses. The unit cell of this new FinFET synapse is composed of two identical FinFETs in series; one serves as control and the other one as storage. This new FinFET synapse exhibits ideal linearity with nearly infinity training cycles (>10^{12}), much lower programming voltage, 0.85V, and faster speed, 2.5ns. It can also analogically increase or decrease the transistor's V_{th} to vary the drain conductance. As far as the analog performance is concerned, it performs excellent linearity and a wide tuning-window (20x) of weight-tuning capability. 1kb synaptic array has also been designed. The spice-simulated results have shown that new FinFET synaptic array can expand the array-size to 64x64, exhibiting 300x of SNR, w.r.t. that of RRAM array. Finally, the training of the neural network based on the proposed FinFET synapse can achieve 97.43% accuracy as high as the GPU one does.

1. Introduction

Deep-neural network (DNN), as one of AI machine learning algorithms, has achieved tremendous progress, including image or speech recognition, gaming, and real-time languages. Recently, many efforts have been paid on using RRAM [1], CBRAM [2], and PCM [3] as analogue memory, implemented in electric synapse devices of DNN. However, resistance-based approach suffered serious variability and nonlinear tuning during the weight update. On the other hand, the charge-storage NVM, such as Floating gate [4] and SONOS [5], shows better linear-tuning capability. But, the programming voltages are too high (6V~20V). Recent report showed that DRAM as the capacitance-based electric synapse becomes a good candidate [6], but has drawback of large size of DRAM capacitance.

In this work, it is of interest for us to present a different strategy by using FinFET as synapse. We will utilize a fully-connected neural network (NN), Fig. 1, to learn and to recognize tested images. The unit cell consists of two identical FinFETs: one is the control transistor and the other is the storage with the gate floating, both together perform the function of a 1T1C, Fig. 1(f), similar to a DRAM. By doing this, a pure logic FinFET can replace conventional large size of DRAM capacitance, which dramatically reduces the layout size of synapses. Therefore, firstly we will elucidate principles of the simple 1T1C architecture. Then, analog tuning characteristics will be evaluated and its performance will be compared with the RRAM array by simulation. Finally, we will do the benchmark on how FinFET-based synapse is superior to the others.

2. Fabrication of FinFET Synapses

The experimental devices used in this work were developed from UMC 14nm FinFET platform. P-channel FinFET was used in the study. Si Fins have been defined before the punch-through stopper implant into the Fin-bottom. The epied SiGe S/D with in-situ p-type impurities have been formed and attached on the channel. After removal of the sacrificial poly-gate deposition on a 0.8nm SiO_2, the high-k and metal gate last process were carried out on the SiO_2 by ALD. Different dimensions of devices with various fin numbers have been prepared.

3. Results and Discussion

A. Operation Principle of the FinFET Synapses: Fig. 2 shows the DC sweep of the new FinFET synapse. To be noted that if the gate is kept floating without any biases, the storage FinFET can be treated as a back-to-back pnp whose characteristics perform a perfect symmetrical bipolar behavior, ①-to-②. When the device is swept backward, steps ③-to-④, we observe a noticeable hysteresis which creates enough ΔV_{th} and on-off ratio. This is similar to the dynamic behavior of DRAM. The physical mechanisms to explain the results in Fig. 3 are divided into *a* to *e* steps. *a* is the equilibrium state; as the reversed V_{ds} increases, *b*, impact ionization happens in the depletion region of channel, the excess electrons are generated and stored in the channel; when the reversed bias drives even higher, *c*, the impact ionization evolves into the avalanche, and a large quantity of electron-hole pairs generated, results in the storage of excess electrons in the channel. When the forward bias is applied, *d*, the stored electrons are generated, in addition to the holes (majority carriers). Finally, after pulling-out electrons, in *e*, only holes are present, and the current falls back to the initial state. The hysteresis disappears. Fig. 4 compares the forward hysteresis (H_{For}) and the reversed one (H_{Rev}). H_{Rev} is larger than H_{For}, which means there exists a best condition of the programming (PGM) energy efficiency. Fig. 5 shows that the higher the PGM reversed voltage is, the lower the energy efficiency becomes. Therefore, the PGM voltage can be set to around -0.8V. Finally, Fig. 6 provides a modeled equation of the stored (residual) charges, which is exponentially dependent on PGM voltage.

B. Analog Characteristics of FinFET Synapse : Fig. 7 shows the cumulative distribution of 1T1S for bit-0 and bit-1, collected by 50 samples. 15x of on/off ratio with a steep distribution has been achieved. Fig. 8 shows the experimental read-out waveforms for bit-1/0, respectively. Fig. 9 is programming transient results of FinFET synapses. By applying -0.85V of PGM voltage and 2.5ns duration pulse, one can successfully program the unit cell. Fig. 10 shows the endurance, which is able to be cycled up to 10^{12} times and provides a 18x of on/off ratio. For the analog properties, the scheme of programming pulses with a ramped pulse, in Fig. 11, was used and a perfect linearity of conductance can be achieved. Also, the tuning window can be enlarged to 20x.

C. Implementation of FinFET Synaptic Array: Here, we use a 4-layer fully-connected neural network (FC-NN), Fig. 1, to learn and to recognize tested images correctly from the database. However, it is important to notice that there is a need to tune the negative or positive values, that is, the G^+ or G^-. In Fig. 1(d), G^+ or G^- can be implemented by FinFET synapses. To evaluate the circuit performance of synaptic array, we use Spice to model the current unit cell (2 FinFETs) as a series connection of 1T+1C, given in Fig. 1(f), for circuit level simulations. Fig. 12 shows a perfect symmetry of linear weight updates which forms a *quasi-isosceles triangle* with almost the same angles at the base. Each up-hill or down-hill needs only 2.5ns. Using 1S1R (1 selector and 1 RRAM) array as the reference, Fig. 13 shows that the SNR of FinFET synaptic array reveals 300x of improvement. Figs. 14 and 15 show the active and standby power consumption of FinFET synapse, compared to that of 1S1R ones. Fig. 16 shows the testing accuracy of 4-layer FC-NN with different delay times. It was found, since the retention time (1-10seconds) of the storage (1C) is much longer than 100μs, it will not show any loss of the accuracy which can reach 97.43% in comparison to ideality (pink colored curve) without charge leak.

In summary, in Table 1, this work demonstrated successfully a FinFET synapse for AI training, which can be fully-integrated in advanced FinFET platform without any extra cost and masks. With a small feature size, $8F^2$, high-programming speed, 2.5ns, low PGM voltage, -0.85V, and 10^{12} cycles of endurance, the FinFET synapse demonstrated a great potential as the unit cell of synaptic array. For the benchmark at the array level, compared to 1S1R cross-bar array, current 1T1C array reveals excellent linear and symmetrical updates, 300 folds of SNR improvement, and much lower active power consumption (~20μW) and standby power (10nW). These results will further strengthen the potentials of FinFET synapses in AI learning and neuromorphic computing.

Acknowledgments This work was supported by the Ministry of Science & Technology under MOST106-2221-E-009-130, Research of Excellence program ***MOST****107-2633-E-009-003*, and *NCTU-Stanford* Dragon-gate program, *MOST 107-2911-I- 009-514*.

References:

[1] W. Wu et al., ***VLSI Tech.***, 2018, p. 103. **[2]** Y. J. Jeong et al., ***APL,*** 173105, 2015. **[3]** S. Ambrogio et al., ***Nature***, p. 60, 2018. **[4]** J. -H. Lee et al., ***VLSI Tech.***, 2018, p.169. **[5]** H. -T. Lue, et al., ***VLSI Tech.***, 2018, p.177. **[6]** Y. Liao et al., ***VLSI Tech.***, 2018, p. 29.

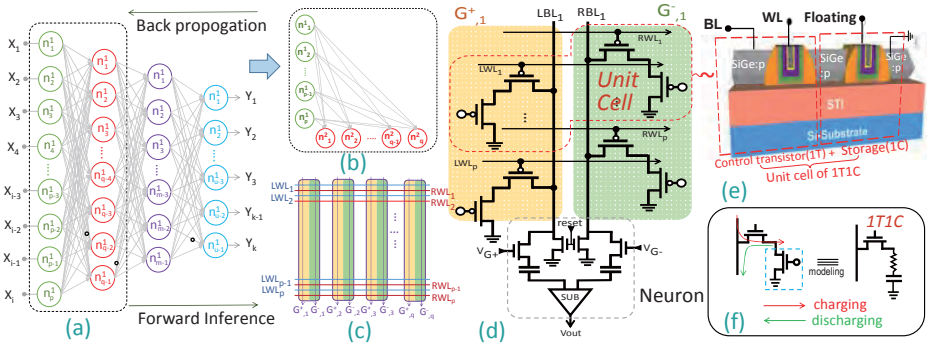

Fig. 1 (a) A 4 layers Fully-connected neural network, **(b)** re-arrangement of the first 2 layers of **(a),** from parallelization into orthogonality with a matrix fashion in **(c)**, **(d)** the zoom-in of a column in **(c)**, **(e)** each unit cell consists of 2 identical FinFETs, and **(f)** two transistors as one unit cell is modeled as *1T1C*.

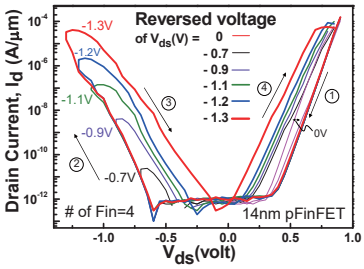

Fig. 2 The forward and backward sweep of FinFET storage which shows a bipolar characteristics from ① to ②. During the backward, ③ to ④, excess charges are generated in the channel and create the hysteresis.

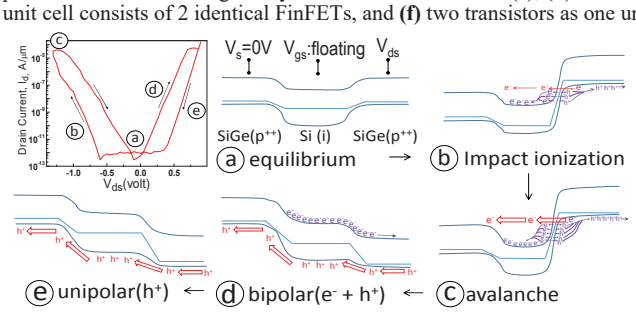

Fig. 3 The physical mechanism of hysteresis: 5 steps are involved in the creation of hysteresis (top-left corner) characteristics of the storage, similar to the DRAM charge-discharge behavior.

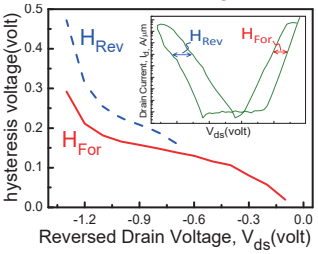

Fig. 4 The comparison of the internal voltages induced by the hysteresis at reverse bias(H_{Rev}) and forward bias (H_{For}), which are defined in the insert.

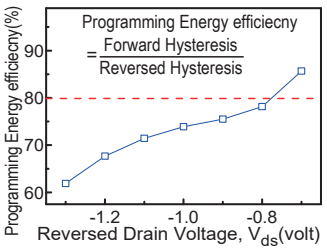

Fig. 5 Calculation of PGM energy efficiency, defined as ratio of forward hysteresis to reversed one, i.e., how many charges gained in the reversed sweep can be used to induce V_{th}-shift in the forward sweep.

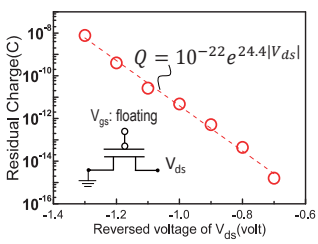

Fig. 6 Modeled (residual) charges in the channel during the reversed sweep as function of the applying voltages.

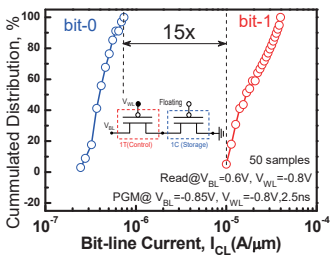

Fig. 7 The statistics of cumulated distributions of bit-0 and bit-1 in FinFET Synapse. 15x of on/off ratio has been achieved.

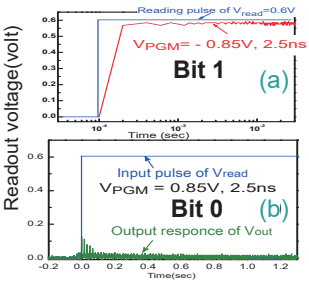

Fig. 8 Real time response of the program and readout pulse of FinFET synapse. **2.5ns** of the speed PGM is demonstrated, for **(a)** bit1and **(b)** bit 0.

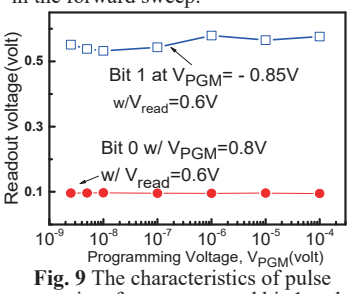

Fig. 9 The characteristics of pulse transient for programmed bit-1 and bit-0 of FinFET synapse. With only -0.85V during 2.5ns, the unit cell can be programmed successfully.

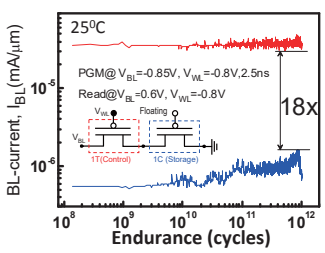

Fig. 10 The endurance of FinFET synapses can be up to 10^{12} cycles while still keep sufficient window.

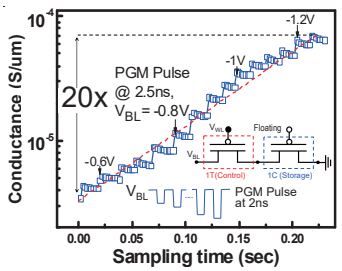

Fig. 11 The experimental conductance of the tuning weight by applying a programming pulse with a varying amplitude, from which a very large tuning window, 20x can be achieved.

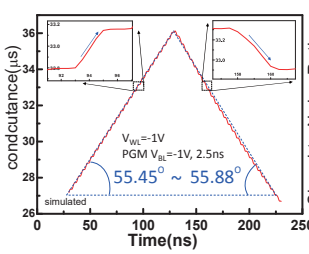

Fig. 12 Simulation results of the conductance in a FinFET synaptic unit cell, showing very linear and symmetrical characteristics with almost the same slope.

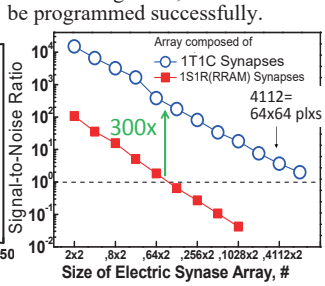

Fig. 13 FinFET synaptic array has 300-fold enhancement of the signal-to-noise ratio (SNR) compared to 1S1R cross-bar one.

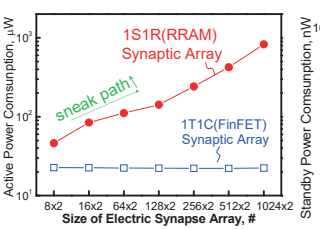

Fig. 14 The active power consumption of FinFET synaptic array is insensitive to the array size, which allows further extension of array size.

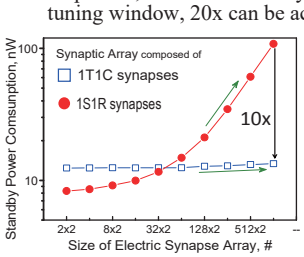

Fig. 15 The standby power consumption of FinFET synaptic array can be kept at a very lower level, ~10nW.

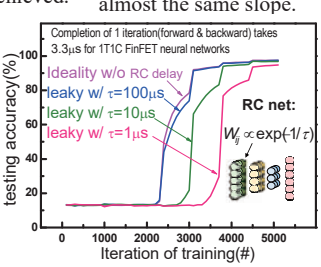

Fig. 16 Testing accuracy of the network in consideration of different RC delay time (τ).

	RRAM [1]	CBRAM [2]	PCM [3]	SONOS [5]	This work
Architecture	1T1R	1R	2R+3T1C	1T	1T1C
On-off ratio	10	10	10	10	20
PGM speed (s)	50n	100n	14n~4.4μ	10μ	2.5n
PGM Voltage(V)	1.5	1.6	1.8	8.5	0.85~1.2
Linearity	⌒	⌒	NA	⌒	⌒
Retention (s)	10	NA	NA	10^5	1~10
Endurance(#)	NA	NA	NA	1K	10^{12}

Table 1 Comparisons of key features in this work with the other reported synaptic devices. Salient features include very high PGM speed and low operation voltage, high linearity etc.