CTT-based Non-Volatile Deep Neural Network Accelerator Design

Yang Xiao*, Wuyu Fan*, Yuan Du, Li Du, and Mau-Chung Frank Chang, *Life Fellow, IEEE*
{yuandu,ldu}@nju.edu.cn

Abstract— **CMOS-compatible Charge-Trap Transistor (CTT) has empowered Non-Volatile Memory (NVM) and Compute-in-Memory (CiM) applications using its threshold voltage (V_{TH}) programmability. In the paper, we characterized the threshold voltage tuning behavior of CTTs and built the simulation model based on silicon data in 28nm CMOS technology. With the established model, we discussed the design methodology and performance evaluation of non-volatile Deep Neural Network (DNN) accelerators. This paper also summarizes the previous work that CTT is used as an analog multiplier for a DNN accelerator. Then, as a new proof-of-concept design, a CTT-based CiM structure was proposed with normal NVM and computing two operation modes.**

***Index Terms*— Charge-Trap Transistor (CTT), Non-Volatile Memory (NVM), Deep Neural Network (DNN) accelerator, Compute-in-Memory (CiM)**

I. Introduction

Machine learning inference using Deep Neural Networks (DNNs) has provided unprecedented capabilities in various artificial intelligence applications. However, one of the dominant problems that limit its pervasive deployment is that neural networks require great computational energy and throughput resources which are insufficient in modern computing platforms. Compute-in-Memory (CiM) has been proven to have the potential for 10x benefits in both metrics because it greatly reduces the communication cost between storage and processor. Recently, Charge-Trap Transistor (CTT) device has been expected to be a new memory device for CiM which was proven as Non-Volatile Memory (NVM) device in [1] using its threshold voltage (V_{TH}) programmability.

In the paper, we characterized the threshold voltage tuning behavior of CTTs and built the simulation model based on silicon data in 28nm CMOS technology. With the established model, a CTT-based CiM structure was proposed with normal NVM and computing modes. The 784-by-784 CTT-based non-volatile DNN accelerator developed in the previous work [2] was also summarized and discussed.

II. CTT Basics

CTT devices are standard NMOS devices, which are fully logic-CMOS-compatible without adding any process complexity or masks. As is widely used in flash memory design, V_{TH} of CTT devices can be changed by charge trapping operation and modulated by the amount of charge trapped in their gate dielectric. The CTT device's threshold voltage can be increased by applying several positive pulses at its gate. Similarly, a few negative pulses fed into the device gate will reduce V_{TH}.

This work was supported in part by the National Natural Science Foundation of China under Grant 62004097 and 62004096, in part by the Natural Science Foundation of Jiangsu Province under Grant BK20200329. Y. Xiao, W. Fan, Y. Du, L. Du are from the School of Electronic Science and Engineering, Nanjing University, Nanjing 210023, China. M.C. F. Chang is with University of California, Los Angeles, USA, 90095. Y. Xiao and W. Fan contribute equally (Corresponding authors: Yuan Du and Li Du).

III. Silicon Measurement and Modeling

A. Measurement Setup

A test chip is fabricated in 28nm CMOS technology to characterize the charge trapping and de-trapping behavior and the programmable V_{TH}. Isolated NMOS devices (thin-oxide, W/L 600nm/28nm, and thick-oxide, W/L 1um/150nm) were tested and V_G vs. V_D curves were achieved. VDD is set at 0.9V for thin-oxide devices, and 1.8V for thick-oxide devices. The positive programming pulse voltage V_G and negative pulse voltage $-V'_G$ are 1.3V and -1.1V for thin-oxide devices and 2.5V and -2.1V for thick-oxide devices. Programming time T_P stands for the total period of the complete pulse sequences.

B. Device Model

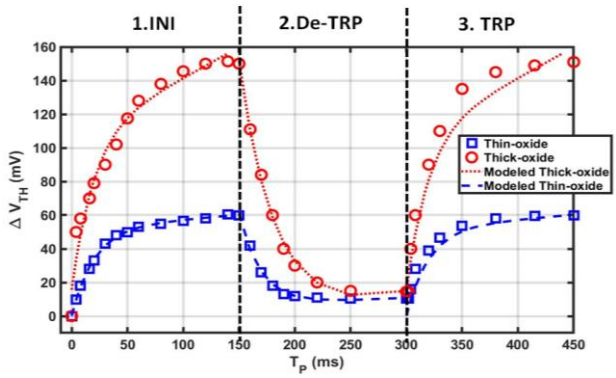

Fig. 1. Threshold voltage changes (ΔV_{TH}) in three phases: 1.INI (Initial trapping); 2. DeTRP (Program for de-trapping); 3. TRP (Program for trapping)

Fig. 1 plots the dynamic behavior of threshold voltage changes (ΔV_{TH}) vs. programming time (T_P) in three phases, INI (Initial trapping), DeTRP (Program for de-trapping), and TRP (Program for trapping). In the process of repeating the DeTRP and TRP phases, the maximal achieved ΔV_{TH} settles at around 150mV and 50mV for thick-oxide devices and thin-oxide devices, respectively.

CTTs were then modeled by replacing the constant V_{TH} in the well-known NMOS transistor current-voltage equation with tunable V_{TH} (T_P), given by:

$$I_D = \begin{cases} \mu_n C_{OX} \frac{W}{L} \{[V_{GS} - V_{TH}(T_p)]V_{DS} - \frac{1}{2}V_{DS}^2\} & ,triode \\ \frac{1}{2}\mu_n C_{OX} \frac{W}{L}[V_{GS} - V_{TH}(T_p)]^2 & ,saturation \end{cases} \quad (1)$$

where V_{TH} (T_P) is the threshold voltage modulated by the programming time T_P, whose empirical relationship is defined by the following equations with R-square value larger than 96%, given by Equation (2) and (3):

$$V_{TH\_thin} = \begin{cases} V_{TH0} + 54.91e^{0.0437T_p} - 4.167e^{-2.263T_p} & ,TRP(T_p < 150ms) \\ V_{TH0} + 3.682e^{-2.467T_p} + 7.956e^{0.1514T_p} & ,DeTRP(150ms < T_p < 300ms) \end{cases} \quad (2)$$

$$V_{TH\_thick} = \begin{cases} V_{TH0} + 126.6e^{0.1093T_p} - 7.955e^{-2.235T_p} & ,TRP(T_p < 150ms) \\ V_{TH0} + 27.31e^{-1.562T_p} + 3.525e^{0.6713T_p} & ,DeTRP(150ms < T_p < 300ms) \end{cases} \quad (3)$$

where V_{TH0} is the initial threshold voltage before the INI phase. This model is then used for further simulations.

IV. Non-Volatile Deep Neural Network Accelerator Design

A. An Analog Neural Network Computing Engine using CTT

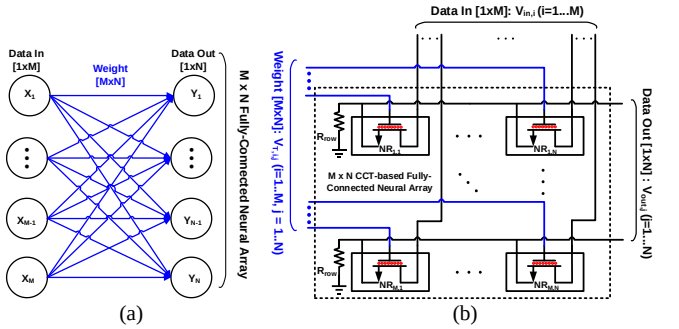

Fig. 2. (a) Fully-connected neural array; (b) CTT multiplication array (Op-amp buffers are not shown)

In [2], we proposed a CTT-based array architecture for an efficient fully-connected layer computation. As shown in Fig. 2 (b), a CTT ($NR_{i,j}$) is a synapse, connecting one input data X_i to one output result Y_j. The device $NR_{i,j}$ stores the synaptic weight ($W_{i,j}$) through threshold voltages V_T. The multiplication operation $W_{i,j} \cdot X_i$ is operated in triode region. The drain-source current I_{DS} of the device approximates a linear function of the product of V_{DS} and V_T as Equation (4) shows:

$$I_{DS}(V_{DS}, V_T) = \frac{1}{2}k_n\frac{W}{L}\left[2V_{DS}(V_{GS} - V_T) - {V_{DS}}^2\right] \approx k_n\frac{W}{L}V_{DS}(V_{GS} - V_T), \text{ when } V_{DS} < V_{GS} - V_T, \quad (4)$$

The summation operation is accomplished by a resistor R_{row} collecting all the drain-source currents from the synapses in the row.

B. CTT-Based CiM structure

As shown in Fig. 4, a CTT-based CiM structure is proposed in this paper, which has two operation modes: normal NVM mode and computing mode.

The basic programming principle of the memory cell follows [1]. However, to maximize the storage density, the proposed architecture uses a single CTT device to store one bit. During writing, the selected BIT CELL's gate is applied positive pulses. The voltage difference between drain and source programs the cell's V_{TH} to low or high. As shown in Fig. 4 (a), by controlling the number of pulses on gate, REF CELL's threshold voltage is programmed to be in the middle of thresholds for logic 1 and logic 0.

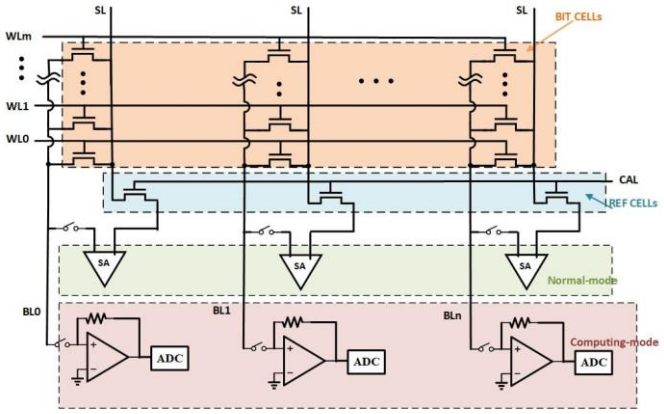

Fig. 3. CTT-based CiM structure with normal NVM and computing two operation modes

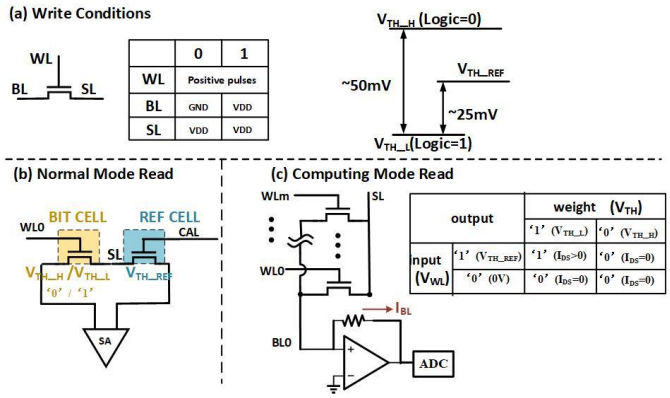

Fig. 4.(a) Write conditions; (b) Normal mode read; (c) Computing mode read.

For normal NVM read, as shown in Fig. 4 (b), due to the V_{TH} difference between BIT CELL and REF CELL, the logic 1 or 0 can be output by a sense amplifier (SA). For computing mode read, by giving different voltages on WLs, drain-source currents (I_{DS}) of BIT CELLs with different weight (V_{TH}) will be different. Multiplication of 1-bit input with 1-bit weight is implemented as shown in Fig. 4 (c). The cell current of the same BL is accumulated to realize the addition operation. The accumulated current of one BL (I_{BL}) is transformed to voltage though the resistor, which is output by an analog-to-digital converter (ADC).

Acknowledgment

The authors would like to thank TSMC for chip fabrication support.

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," *IEEE Electron Device Letters*, vol. 38, no. 1, pp. 44–47, Jan 2017.

[2] 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*, vol. 38, no. 10, pp. 1811-1819, Oct. 2019.