Title
Characterization of the Charge-Trap Transistor for Analog In-Memory Computing

Author
Qiao, Siyun

Publication Date
2022

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Characterization of the Charge-Trap Transistor for Analog In-Memory Computing

A thesis submitted in partial satisfaction of the requirements

for the degree of Master of Science

in Electrical and Computer Engineering

by

Siyun Qiao

2022

ABSTRACT OF THE THESIS

Characterization of the Charge-Trap Transistor

for Analog In-Memory Computing

by

Siyun Qiao

Master of Science in Electrical and Computer Engineering

University of California, Los Angeles, 2022

Professor Subramanian Srikantes Iyer, Chair

Charge-trap transistor (CTT) is a novel non-volatile memory (NVM) technology suitable for both digital and analog applications. In this thesis, we focus on the optimization of CTT to be used as an analog NVM memory element for vector-matrix multiplication for artificial neural network inference. Three important aspects of CTT operation are identified and evaluated. First, we investigate noise-induced fluctuation during read operation both before and after a programming event. Results show that fluctuation of CTT has an insignificant impact on its operation. Second, data encoding precision is studied and characterized. We develop a new

programming protocol to improve encoding accuracy and compare it with the previous scheme. Furthermore, we focus on characterizing the long-term retention of CTT devices. Experiments are performed to characterize retention at both room temperature (RT) and 85℃ conditions. We then optimize the programming protocol to improve retention characteristics.

The thesis of Siyun Qiao is approved.

Sudhakar Pamarti

M. C. Frank Chang

Clarice D. Aiello

Subramanian Srikantes Iyer, Committee Chair

University of California, Los Angeles

2022

To my parents

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION 1

1.1 Motivation of the Work 1

1.2 Fundamentals of Charge-Trap Transistor 3

1.3 Charge-Trap Transistor for In-memory Computing 5

1.4 Organization of this Work..... 7

CHAPTER 2: CHARACTERIZATION OF NOISE-INDUCED VARIATION 8

2.1 Overview 8

2.2 Origins of Noise 9

2.3 Design of Experiments 10

2.3.1 Impact of Programming on Device Noise 10

2.3.2 Impact of Time of Measurements on Device Noise 11

2.4 Results and Discussion 12

CHAPTER 3: CHARACTERIZATION OF PROGRAMMING ACCURACY 14

3.1 Overview 14

3.2 Previous Work 16

3.3 New Programming Scheme with Fine-tune Capability 18

3.4 Trade-off Between Programming Accuracy and Speed 20

3.5 Distribution of States After Programming as a Function of Target Levels 22

CHAPTER 4: CHARACTERIZATION OF DEVICE RETENTION 26

4.1 Overview 26

4.2 CTT Retention Characteristics at Room Temperature 27

4.3 Impact of Programming Voltage on CTT Retention at High Temperature 30

4.4 CTT Retention Characteristics at High Temperature …… 33

4.5 Limitations and Future Plans …… 35

CHAPTER 5: SUMMARY …… 36

LIST OF FIGURES

Figure 1-1: (a) the structure of a single neuron; (b) the function of the neuron is modeled by an activation function; (c) the connection between two neurons is called a synapse; (d) artificial neural network that models the basic structure of the brain ……...…….. 2

Figure 1-2: Schematic representation of GF 22FDX technology …… 3

Figure 1-3: (a) PRG operation; (b) ERS operation …….. 4

Figure 1-4: CTT-based analog in-memory computing scheme …… 5

Figure 3-1: Previous CTT programming scheme ……...……. 16

Figure 3-2: CTT error distribution assuming consistent distribution regardless of target states …… 17

Figure 3-3: New programming scheme with fine-tune capability ……... 19

Figure 3-4: Write-verify demonstrated with 3% accuracy. 7 200us-long programming pulses and 3 1. 6ms -long erase pulses are applied. Erase pulses exhibit smaller current steps compared to programming ……... 19

Figure 3-5: A total of 480 devices are programmed to 6 different target levels (100,200,…,600nA). Each device is programmed to 3% within the target …… 23

Figure 3-6: Neighboring cell half-selection demonstration during PRG operation ……..... 24

Figure 3-7: Data points of mean drift and standard deviation of normal distribution curves for each target level. Quadratic curves are fitted to the data points …….. 25

Figure 3-8: Modeled ENOBs per CTT based on programming distributions obtained in fig. 3-5 and fig. 3-7. Fitted mean drift and standard deviations are used ……..... 25

Figure 4-1: (a) 20hr continuous current measurements of the CTT programmed with only PRG pulses; (b) 20hr continuous current measurements of the CTT programmed with both PRG and ERS pulses …….. 27

Figure 4-2: 18hr continuous measurements on drain current after voltage burn-in ……... 28

Figure 4-3: Device retention at room temperature ……..…. 29

Figure 4-4: Normalized group mean drifts and standard deviations with respect to the target level

…….…… 32

Figure 4-5: (a) Cell distribution after 20 hours at 85℃; (b) Cell distribution after 50 hours at 85℃

……..… 33

Figure 4-6: Whisker-box plot for T=0, T=20hr, and T=50hr distributions ……..… 34

Figure 4-7: (a) Measured and modeled mean drifts right after programming, after 20 hours at 85℃, and after 50 hours at 85℃. (b) Measured and modeled standard deviation right after programming, after 20 hours at 85℃, and after 50 hours at 85℃ …….. 34

LIST OF TABLES

Table 2-1: DOE to investigate the impact of programming on device noise level ……. 11

Table 2-2: DOE for impact of time of measurements ……. 11

Table 2-3: Results of fluctuation of different programming levels ……. 13

Table 2-4: Results of fluctuation of different time of measurements …….. 13

Table 3-1: Results of experiment one with target accuracy set to 1% ……. 21

Table 3-2: Results of experiment two with target accuracy set to 5% …… 21

Table 4-1: Summary of the experiments with results …….. 31

ACKNOWLEDGEMENT

First of all, I would like to express my sincere appreciation to my academic advisor, Prof. Subramanian S. Iyer. Being able to work in Prof. Iyer's lab and learn under his guidance gives me the opportunity to see the wonderfulness of research. He provides me with critical guidance in the project but also allows me with much freedom of exploring the subject by myself which is something I am truly thankful for.

I would also like to thank Prof. Sudhakar Pamarti profoundly for his patient guidance throughout the course of my M.S. project. The way that Prof. Pamarti always critically looks at a subject has brought me with new perspectives of thinking which is extremely important in research.

I would like to extend my sincere appreciation to my colleagues and lab mates. I started my M.S. study during COVID times during which meeting new people and learning from peers were made much more difficult. Fortunately, being able to work in the lab alleviated this issue. Everyday discussions and chats with my lab mates have not only helped with my research but provided profound emotional support. Specifically, I would like to thank my colleagues, Steven Moran, Dhruv Srinivas, Sepideh Nouri, Dr. Zhe (Frank) Wan, and Dr. Faraz Khan for all the support they have given to the project. I would also like to thank my lab mates, Dr. Yu-Tao Yang, Krutikesh Sahoo, Haoxiang Ren, Guangqi Ouyang, Henry Sun, Golam Sabbir, Ankit Kuchhangi, and Vineeth Harish.

Last but not least, I would like to thank my parents, Hongsheng Wang and Huaxiang Qiao, for their unconditional support. My study would have not been made possible without them. I would also like to thank my girlfriend, Fengchen Gong, for being an awesome companion.

Chapter 1

Introduction

1.1 Motivation of the Work

Over the last few decades, the development of artificial intelligence has been in the spotlight for almost every application. In particular, machine learning (ML), a heavily data driven method for analytical model construction, attracts the most attention of researchers and developers around the globe [1-2]. As one of the major branches of ML algorithms, artificial neural networks (ANN) are well-known for their excellent performance in tasks such as image recognition, computer vision, voice recognition, natural language processing, etc. Compared to conventional computer algorithms which are primarily based on mathematical induction and derivation, ANNs mimic the structure the human brain, where the function of a brain neuron is modeled by a nonlinear activation function and the synapse between two different neurons by a value, or a weight [3]. Figure 1 (adapted from [4]) can serve as a demonstration of ANN.

ANN leverages the fact that data processing is done near where the weights are stored in the model. This means computation happens near or within the memory. However, most modern computers feature the von Neumann architecture with separate processing and memory modules. Every time a calculation needs to be performed the processing module sends an instruction to the memory module to fetch the desired data. After the calculation is done, results need to be returned to the memory and stored. This results in significant latency and power consumption when it comes to ANN inferencing and is believed not to be the way the human brain processes information [3].

To realize ANN in hardware, it is of great interest to engineer a system that somewhat resembles the way human brains process information and is capable of doing computation inside the memory itself. This type of system can be very advantageous to be applied to applications that require high throughput, low latency, and low power consumption. At the Center for Heterogeneous Integration and Performance Scaling (CHIPS) at UCLA, a technology called charge trap transistor (CTT) is being developed to be the fundamental building block of such systems.

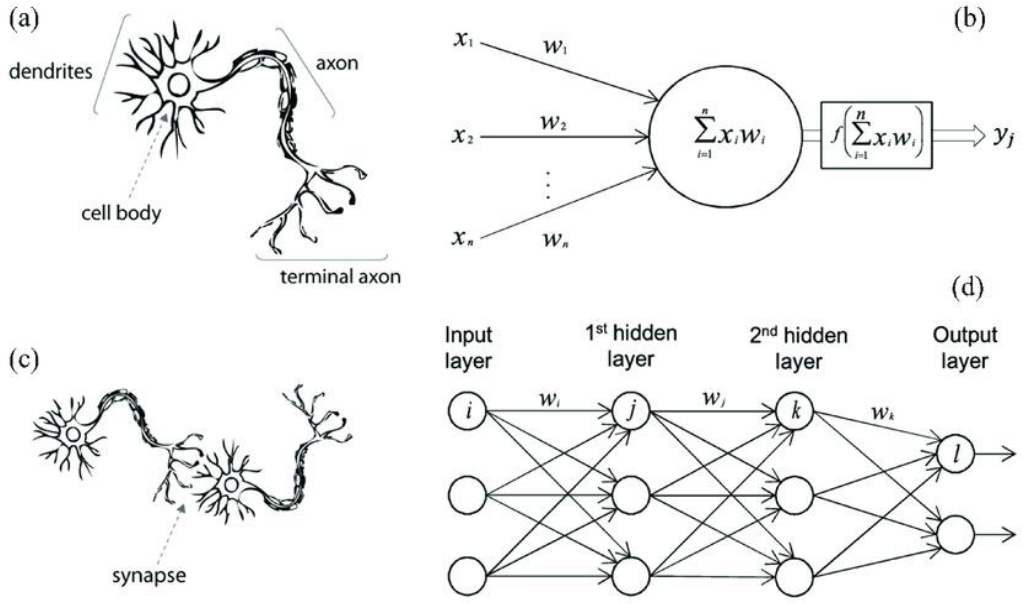


Fig.1-1: (a) the structure of a single neuron; (b) the function of the neuron is modeled by an activation function; (c) the connection between two neurons is called a synapse; (d) artificial neural network that models the basic structure of the brain. [4]

1.2 Fundamentals of Charge-Trap Transistor

Charge trap transistor (CTT) is a CMOS approach that takes advantage of the charge trapping properties of the inherent vacancies in the HfO_x gate dielectric. The charge trapping effects can be enhanced through a process called self-heating, by which the large drain current-induced heat raises device ambient temperature and subsequently creates more defects in the gate dielectric. Charge carriers are prone to be trapped in these defects which in turn alters the threshold voltage.

The technology used in this work as a CTT device is GlobalFoundries 22FDX, which is based on the fully depleted silicon on insulator (FD-SOI) structure and adopts HfO_2 as the high-k gate oxide layer. Figure 2 shows the schematic of such a device.

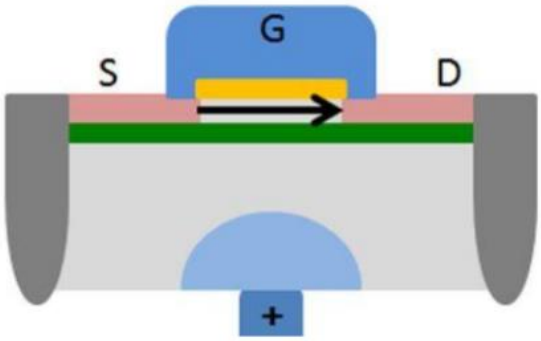
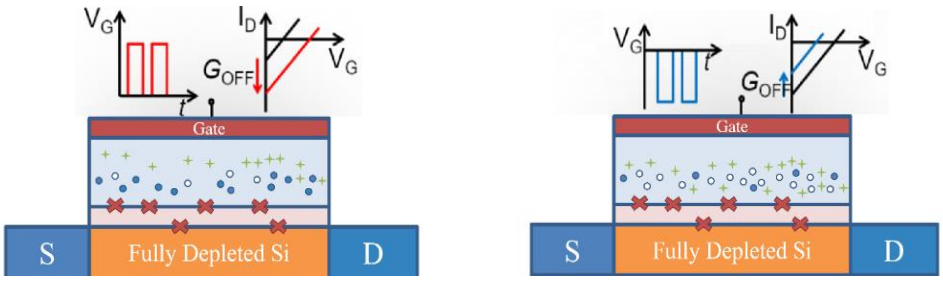


Fig.1-2: Schematic representation of GF 22FDX technology

Threshold voltage of a CTT device can be both increased and decreased. To increase the threshold voltage, a program (PRG) operation is performed, where high drain-to-source and gate-to-source voltage bias conditions are applied. Resulting high channel current generates enough heat to raise the ambient temperature such that charge carriers are more likely to be trapped in the generated defects compared to nominal bias conditions. On the other

hand, an erase (ERS) operation is performed to reduce the threshold voltage. Negative gate-to-source voltage bias is applied while keeping source and drain grounded. The reverse electric field partially erases the trapped charge. Figure 3 shows the schematics for both PRG and ERS operations.



(a) (b)

Fig.1-3: (a) PRG operation; (b) ERS operation

1.3 Charge-Trap Transistor for Analog In-memory Computing

The nature of charge trapping dictates that CTT can be used as an analog memory device. The fundamental minimum resolution is given by the amount of threshold voltage change of one charge carrier trapped or de-trapped. In practice, however, it is unlikely to reach this level of resolution. This is similar to modern Flash memory, where a single memory device can have multiple states instead of just “0” or “1” [5-7].

Operation of ANNs heavily rely on multiplication and accumulation, or MAC. An input to a neuron is first multiplied by the weight on its connection to the next neuron, then added by the products of all the other inputs and their corresponding connections to the same next neuron, as indicated in fig.1 (b). This MAC operation can be fully realized by CTT-based array structure.

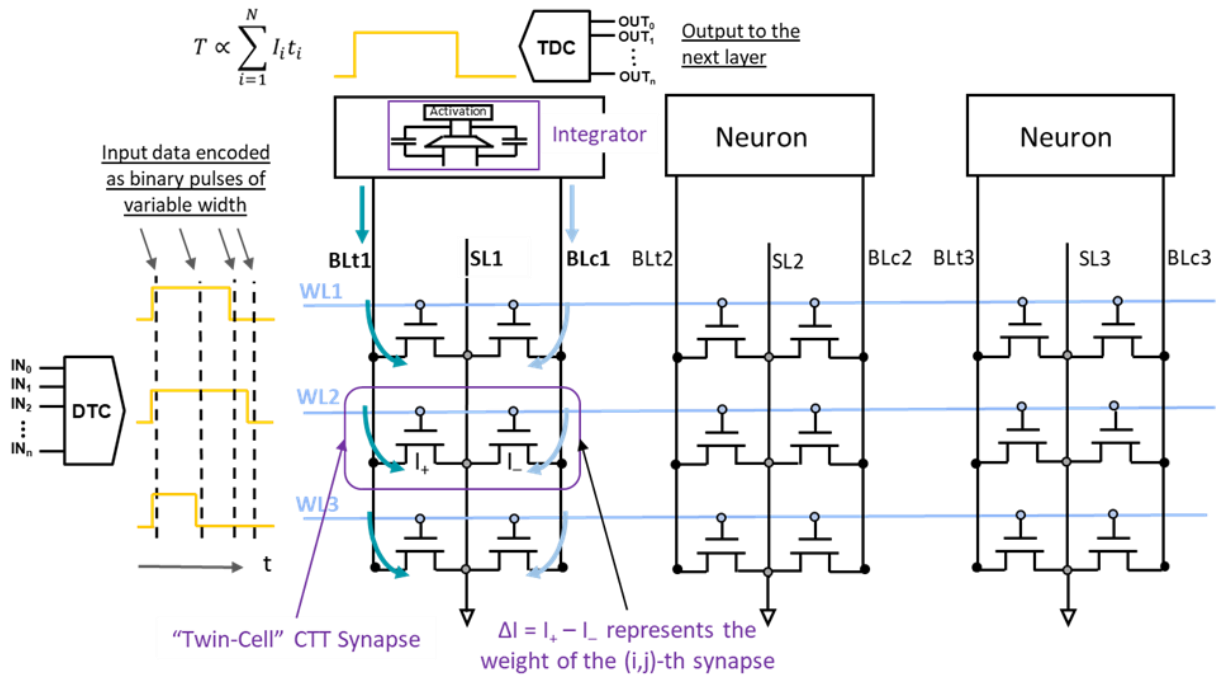


Fig.1-4: CTT-based analog in-memory computing scheme

After data is encoded into the CTT in the form of threshold voltage or, in turn, conductance, computation can be done by leveraging the physics laws. First of all, Eq.1 exhibits that multiplication can be performed using Ohm's law, where inputs are represented by the time period the gate and drain voltages (which are held constant) are applied and weights by the currents (which are the product of voltage and conductance). Second, all the charge that is passed through the line of connection is summed and accumulated onto a capacitor, as given by Eq.2. This process is again illustrated by fig.4, where the CTT-based analog in-memory computing scheme is shown. Note that the differential twin-cell structure is adopted in this case to represent both positive and negative numbers.

$$Q_{out,sub} = t_{input} \times I_{weight} = t_{input} \times (V_{constant} \times G_{channel}) \quad \ldots \qquad \text{Eq.1}$$

$$Q_{out,total} = Q_{out,sub1} + Q_{out,sub2} + \ldots + Q_{out,subN} \ldots \qquad \text{Eq.2}$$

1.4 Organization of the Work

This work aims to provide a thorough investigation of the CTT as a novel non-volatile memory (NVM) device for analog compute-in-memory applications. Specifically, three aspects of the CTT are discussed in detail: noise-induced variation, programming accuracy, and retention characteristics.

The rest of this paper is organized as follows. Chapter 2 discusses noise-induced fluctuation in CTT. Chapter 3 focuses on studying programming accuracy. Chapter 4 mainly characterizes the CTT retention at room temperature and high temperature environments.

Chapter 2

Characterization of Noise-induced Variation

2.1 Overview

As discussed in Chapter 1, weights of an ANN are encoded as channel conductance of CTT devices by modifying threshold voltage, which are then represented by drain current readouts with constant drain and gate voltage biases. Therefore, it is desirable that current outputs remain stable during ‘read’ operation and are immune to noise. There are many factors that can influence the stability of the current output of a device. First of all, it is unclear if a virgin device (devices that have not been unprogrammed) would have less current fluctuation compared to a programmed device, considering that the PRG operation physically introduces defects to the gate dielectric thus making it more susceptible to noise. In addition, the extent to which the device is programmed can also bring an impact. On the other hand, the time at which measurements are made can be another variable. For example, current read right after versus sometime after PRG may result in different fluctuation levels. Systematic investigation is needed to study the impact of each variable.

2.2 Origins of Noise

Two most widely recognized sources of noise are low-frequency noise (LFN), also commonly referred to as 1/f noise or flicker noise, and random telegraph noise (RTN). Publications in [8-9] discuss the potential mechanism of LFN and suggest that LFN is largely related to the Si-SiO_2 traps. In the case of RTN, it is more likely to occur when defects in the bulk of the oxide layer constantly trap and de-trap charge carriers [10-12]. In the structure of GF 22FDX devices, the thin SiO_2 interface layer can introduce interface traps near the Si channel, where the bulk HfO_2 layer can introduce trapping sites for RTN behaviors.

2.3 Design of Experiments

2.3.1 Impact of Programming on Device Noise

To study how devices of different programming levels can change the output current fluctuation, three groups of devices are selected with each group having three devices. Table 1 summarizes this experiment. All devices are read out with drain and gate voltage at 200mV and programmed using a technique called pulsed gate voltage ramp sweep (PVRS) to speed up the PRG process while enabling finer control [13]. Gate and drain voltage pulses are controlled at ~500us. Each device is sampled at 1Hz for 5 minutes. Current fluctuation is measured by finding the average and standard deviation of the sampled current values and dividing the two. Measurements are done immediately after programming ends. The only variable that is held different among the three groups is ΔI_{read}. Devices in group 1 are unprogrammed, so ΔI_{read} = 0nA. ΔI_{read} of devices in group 2 is controlled to be ~200nA, while group 3 has ~700nA ΔI_{read}. It is worth noting that programming stops when ΔI_{read} of each device is past its target value. Since no fine-tune programming techniques are used in this case, it is difficult to achieve the exact desired ΔI_{read}. Nevertheless, this should not affect the outcome of this experiment as the differences between three groups are large enough to ignore programming error.

	# of Devices	PRG Conditions	I_{read} Conditions	**Levels of PRG**	Time of Measurements	Measurements of Fluctuation
Group 1	3	PRG: V_D= 1.2V, V_G= 1.5 - 2.5V, ΔV_G= 50mV t_{PULSE}= 500us, PVRS	V_D=200mV V_G=200mV	**Virgin**	N/A	stdev / mean
Group 2	3	PRG: V_D= 1.2V, V_G= 1.5 - 2.5V, ΔV_G= 50mV t_{PULSE}= 500us, PVRS	V_D=200mV V_G=200mV	**ΔI_{read} ~ 200nA**	Right After PRG	stdev / mean
Group 3	3	PRG: V_D= 1.2V, V_G= 1.5 - 2.5V, ΔV_G= 50mV t_{PULSE}= 500us, PVRS	V_D=200mV V_G=200mV	**ΔI_{read} ~ 700nA**	Right After PRG	stdev / mean

Table 2-1: DOE to investigate the impact of programming on device noise level

2.3.2 Impact of Time of Measurements on Device Noise

To study the effect of time of measurements on device noise, previous DOE can again be utilized by continuing measuring the same devices at different time points. Specifically, 200 hours and 400 hours after the devices are programmed are chosen for measurements. The outcome of this experiment will be compared to that of table 1, namely the fluctuation level right after programming, to help investigate whether device noise is a function of time. Table 2 summarizes the DOE of this experiment.

	# of Devices	PRG Conditions	I_{read} Conditions	Levels of PRG	**Time of Measurements**	Measurements of Fluctuation
Group 1	3	PRG: V_D= 1.2V, V_G= 1.5 - 2.5V, ΔV_G= 50mV t_{PULSE}= 500us, PVRS	V_D=200mV V_G=200mV	Virgin	**N/A**	stdev / mean
Group 2	3	PRG: V_D= 1.2V, V_G= 1.5 - 2.5V, ΔV_G= 50mV t_{PULSE}= 500us, PVRS	V_D=200mV V_G=200mV	ΔI_{read} ~ 200nA	**200hr After PRG 400hr After PRG**	stdev / mean
Group 3	3	PRG: V_D= 1.2V, V_G= 1.5 - 2.5V, ΔV_G= 50mV t_{PULSE}= 500us, PVRS	V_D=200mV V_G=200mV	ΔI_{read} ~ 700nA	**200hr After PRG 400hr After PRG**	stdev / mean

Table 2-2: DOE for impact of time of measurements

2.4 Results and Discussion

Experiments are conducted and the results are shown in the following tables. The average values of standard deviation-over-average are calculated for each experimental group. As can be seen in table 3, all three groups have very similar levels of fluctuation despite different levels of initial programming. In addition, there is no clear trend that shows more device programming leads to more/less fluctuation. This indicates that device programming is irrelevant to noise-induced current output variation.

In addition, table 4 shows that current fluctuation is unlikely to be a function of time after programming ends. There is no clear correlation that can be identified to demonstrate that noise-induced fluctuation changes as time increases.

Results from these two sets of experiments are particularly encouraging when it comes to applying CTT as a non-volatile memory device for weight storage for analog in-memory computing applications, due to the fact that stable current outputs are very much desired for minimal impact on system noise. Moreover, the invariance of noise over time makes system modeling possible because of the predictability of device current output. Similar levels of fluctuation are reported for Flash memory, which are considered as a mature NVM technology [14].

	# of Devices	PRG Conditions	I_{read} Conditions	**Levels of PRG**	Time of Measurements	**Measurements of Fluctuation**
Group 1	3	PRG: V_D= 1.2V, V_G= 1.5 - 2.5V, ΔV_G= 50mV t_{PULSE}= 500us, PVRS	V_D=200mV V_G=200mV	**Virgin**	N/A	**stdev / mean = 0.76%**
Group 2	3	PRG: V_D= 1.2V, V_G= 1.5 - 2.5V, ΔV_G= 50mV t_{PULSE}= 500us, PVRS	V_D=200mV V_G=200mV	**ΔI_{read} ~ 200nA**	Right After PRG	**stdev / mean = 0.68%**
Group 3	3	PRG: V_D= 1.2V, V_G= 1.5 - 2.5V, ΔV_G= 50mV t_{PULSE}= 500us, PVRS	V_D=200mV V_G=200mV	**ΔI_{read} ~ 700nA**	Right After PRG	**stdev / mean = 0.79%**

Table 2-3: Results of fluctuation of different programming levels

	# of Devices	PRG Conditions	I_{read} Conditions	Levels of PRG	**Time of Measurements**	**Measurements of Fluctuation**
Group 1	3	PRG: V_D= 1.2V, V_G= 1.5 - 2.5V, ΔV_G= 50mV t_{PULSE}= 500us, PVRS	V_D=200mV V_G=200mV	Virgin	**N/A**	**stdev / mean = 0.72% (200hr) 0.73% (400hr)**
Group 2	3	PRG: V_D= 1.2V, V_G= 1.5 - 2.5V, ΔV_G= 50mV t_{PULSE}= 500us, PVRS	V_D=200mV V_G=200mV	ΔI_{read} ~ 200nA	**200hr After PRG 400hr After PRG**	**stdev / mean = 0.82% (200hr) 0.75% (400hr)**
Group 3	3	PRG: V_D= 1.2V, V_G= 1.5 - 2.5V, ΔV_G= 50mV t_{PULSE}= 500us, PVRS	V_D=200mV V_G=200mV	ΔI_{read} ~ 700nA	**200hr After PRG 400hr After PRG**	**stdev / mean = 0.77% (200hr) 0.81% (400hr)**

Table 2-4: Results of fluctuation of different time of measurements

Chapter 3

Characterization of Programming Accuracy

3.1 Overview

Data encoding methodology is among the most important topics of study and research for any multibit/analog memory technology. Contrary to traditional digital memory where only two states, namely “0” and “1”, can be represented, multibit memory can have a memory window that is large enough to include multiple different states within one single cell. Analog memory can even have an infinite number of states in theory, but in practice it is usually limited by the accuracy at which the data can be encoded.

A typical example of multibit memory is Flash memory. In the early 2000s, the introduction of Mirrorbit technology by Spansion became one of the first 2-bit-per-cell NOR Flash memory which remains to be among the most attractive NOR Flash technologies even today [15]. It differs from the conventional floating gate based Flash memory in that it uses a nitride layer for charge storage. The long channel structure and the channel hot electron (CHE) injection method used for programming enables the device to be programmed individually on both its source end and drain end. This effectively gives the device the ability to hold four individual states, or 2 bits equivalently. More recent developments of Flash involve triple-level cell (TLC) [7] and quadruple-level cell (QLC) [6], which correspond to 3-bit-per-cell or 8 states, or 4-bit-per-cell or 16 states, respectively.

The key to having such a large number of states within a single cell lies in the ability to encode each state accurately. The failure to do so can cause significant overlap of different states and thus bit error [16]. This becomes even more important when it comes to analog memory because of its continuous nature of states. In this chapter, a data encoding scheme for CTT with fine-tuning capability is introduced. Average numbers of programming pulses for different accuracy requirements are obtained. Corresponding programming time per cell is estimated.

3.2 Previous Work

Previously developed CTT data encoding scheme is shown in fig.5. Since the twin-cell structure is adopted to represent signed data, the differential nature of this is used for data encoding. Essentially, the difference between two cells' current is the encoded data, so if this difference is larger than the target, it is suppressed by programming the "true" device; if this difference is less than the target, it is increased by programming the "complementary" device. This scheme does not use ERS pulses to fine-tune the state of the cell. However, due to the fact that PRG operation involves hot carrier injection to a certain degree, it is inherently harder to control compared to ERS. In addition, being able to target specific states in a device accurately makes it easier for large scale characterization, which can be beneficial for system level analysis. Therefore, a new data encoding scheme with fine-tune capabilities is desired.

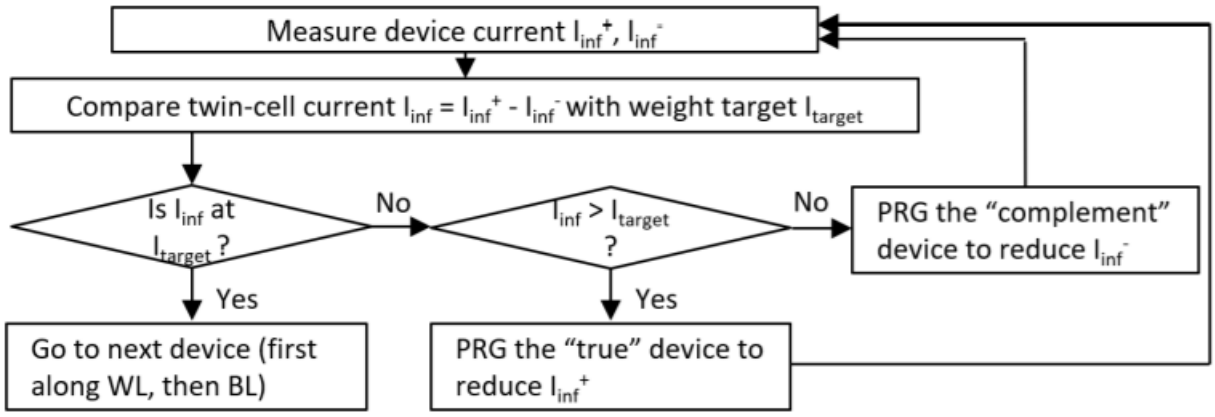


Fig. 3-1: Previous CTT programming scheme

In terms of characterization of programming accuracy, an assumption that there is a normal error distribution model that can be applied to all devices regardless of target state has

been made. In other words, devices programmed to 500nA follow the same error distribution compared to the ones programmed to 5nA. Based on this assumption, experiments have been conducted to collect data for model construction. Fig. 6 (obtained from [8]) shows the distribution of device state versus target. However, because this experiment intentionally ignores the variability of device states, it is unclear whether error distribution can also be a function of device states. Therefore, it is valuable to further investigate this issue to potentially build a more accurate error distribution model.

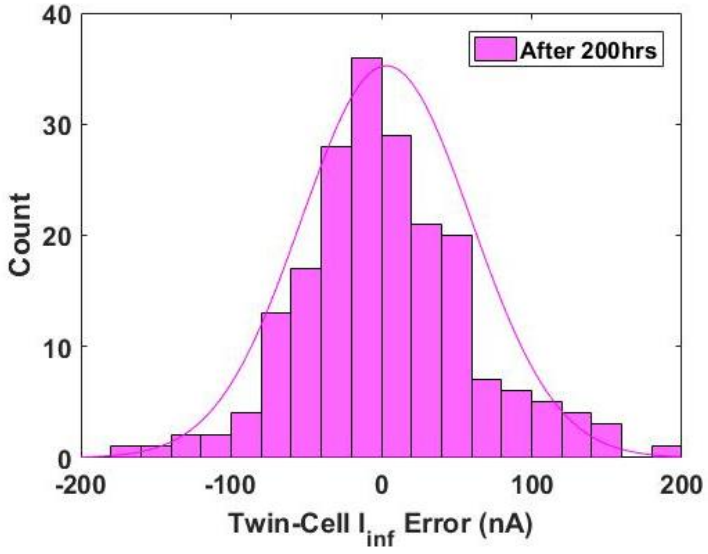


Fig. 3-2: CTT error distribution assuming consistent distribution regardless of target states

3.3 New Programming Scheme with Fine-tune Capability

To develop a new data encoding scheme with fine-tune capability, a combination of PRG and ERS pulses need to be used. First, only program pulses are applied to the target device until the current readout drops below the target level by a certain amount. This is given by how much the device is desired to be over-programmed to ensure enough room for fine-tuning. Model parameters are constantly updated depending on the state of programming. Then, the fine-tuning process begins with a sequence of ERS and PRG pulses. The programming event would be terminated when the current readout sits within the upper and lower bounds of the target. The scheme is shown in fig.7. Fig.8 provides an example of programming an as-fabricated CTT device following the scheme with the target level set to 200nA. The nonlinear and relatively large drop in device current during consecutive PRG pulses indicates that it is difficult to program the device to a precise target using only PRG pulses. During the fine-tuning stage, device current changes at a much slower rate compared to the previous case, which is very beneficial for fine-tuning. The combination of both makes it possible to accelerate the device programming process without compromising accuracy. However, it is inevitable that there is a trade-off between accuracy and speed. Higher programming accuracy (e.g., error is <3%) would statistically require more fine-tune steps which increases the total programming time.

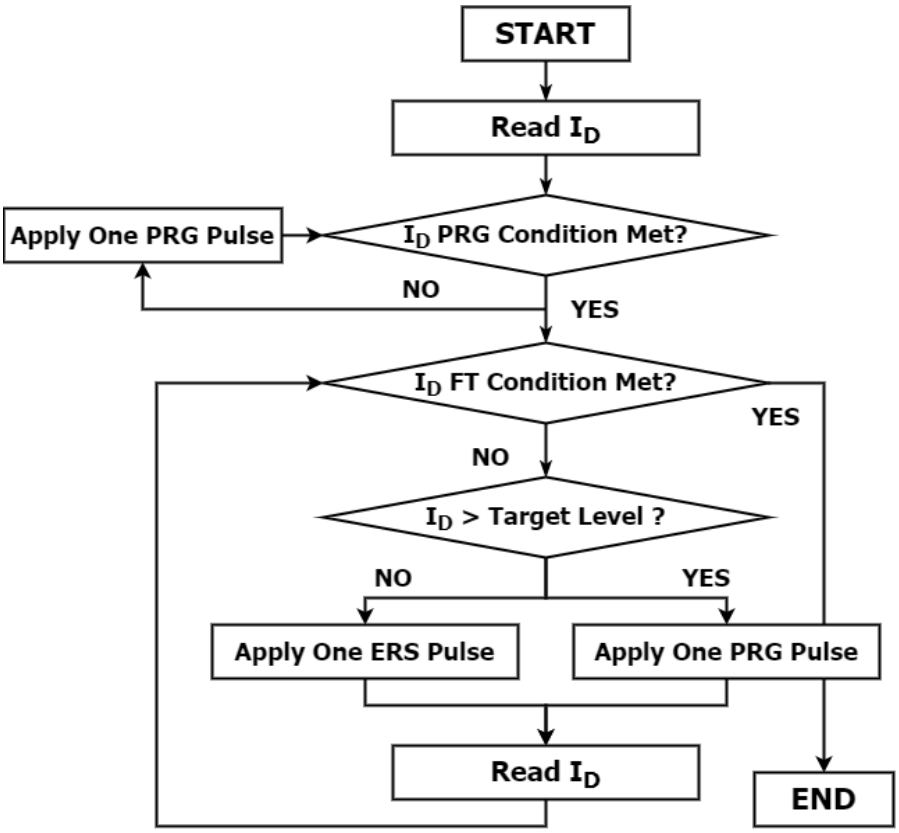


Fig. 3-3: New programming scheme with fine-tune capability

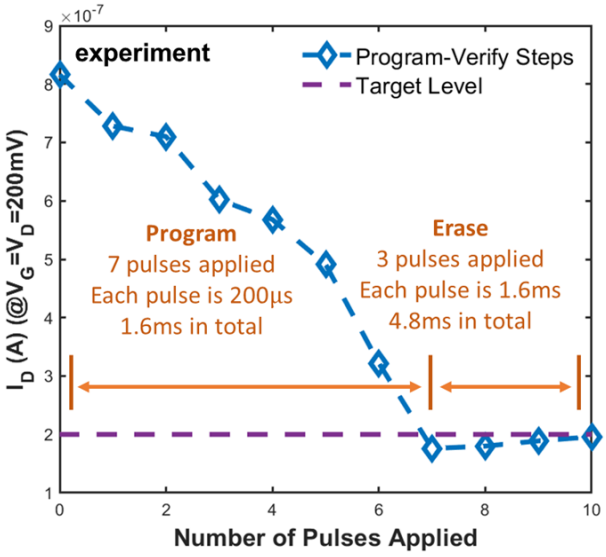


Fig. 3-4: Write-verify demonstrated with 3% accuracy. 7 200us-long programming pulses and 3

1. 6ms -long erase pulses are applied. Erase pulses exhibit smaller current steps compared to

programming

3.4 Trade-off Between Programming Accuracy and Speed

To test the new data encoding scheme, two experiments are planned and conducted. For each experiment, a total of 30 devices are randomly selected grouped into 6 sets, with each set containing 5 devices to be programmed to a specific target level. For PRG, a constant 1.2V is applied to the drain, and the gate voltage is ramped from 1.5V with an increment of 50mV with source grounded. For ERS, the gate voltage is ramped from -1.5V with an increment of -50mV with drain and source being grounded. Both PRG and ERS pulses are 500us long. All devices in the first experiment have 1% programming accuracy, meaning programming stops when the current readout converges within ±1% of the target. All devices in the second experiment have 5% accuracy.

Results from experiment one and two are shown in table 5 and table 6, respectively. In both experiments, with lower target levels, more pulses are needed to reach the preset accuracy. This is expected since the lower the target level, the more charge needs to be trapped in the dielectric, therefore the more programming pulses. By comparing experiment one and two, it can be observed that lower accuracy (5%) requires fewer number of pulses to reach convergence. This is, again, within the expectation because lower programming accuracy requires less fine-tune process to achieve the target range, thus shorter programming time or faster speed.

	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6
Target (nA)	**600**	**400**	**200**	**100**	**50**	**20**
# of Devices	**5**	**5**	**5**	**5**	**5**	**5**
Target Accu.	**1%**	**1%**	**1%**	**1%**	**1%**	**1%**
Avg. # of Pulse	**13.4**	**17**	**25.4**	**22.4**	**23.2**	**24.2**
Avg. Time (ms)	**6.7**	**8.5**	**12.7**	**11.2**	**11.6**	**12.1**

Table 3-1: Results of experiment one with target accuracy set to 1%

	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6
Target (nA)	**600**	**400**	**200**	**100**	**50**	**20**
# of Devices	**5**	**5**	**5**	**5**	**5**	**5**
Target Accu.	**5%**	**5%**	**5%**	**5%**	**5%**	**5%**
Avg. # of Pulse	**6.2**	**7.4**	**9.6**	**11.4**	**13.2**	**14.6**
Avg. Time (ms)	**3.1**	**3.7**	**4.8**	**5.7**	**6.6**	**7.3**

Table 3-2: Results of experiment two with target accuracy set to 5%

3.5 Distribution of States After Programming as a Function of Target Levels

As mentioned earlier in this chapter, the assumption made previously suggests that CTT state distribution after programming follows a uniform distribution model and is irrelevant to target states. In order to verify if this is a valid assumption, a total of 480 devices have been randomly assigned to six different target state levels (100, 200, ..., 600nA) with each level having 80 devices. Device programming is done following the new scheme discussed in 3.3, with 2V drain bias and gate voltage ramping from 1.5V to 2.5V for PRG pulses (200μs/pulse), and with 1.4V source/drain bias and gate ramping from -0.4V to -1.4V for ERS pulses(1.6ms/pulse). Programming stops when measured current is within 3% error of the target. After all cells are programmed, a voltage burn-in is performed (discussed in more detail in Chapter 4). The distribution of states and the fitted normal distribution curves are shown in fig. 9.

From fig. 9, it can be observed that 1) state distribution is inconsistent with six different targets, 2) distributions are wider at higher target levels, and 3) although all devices are programmed such that each individual cell should have less than 3% error, the distribution exhibits that larger error terms exist.

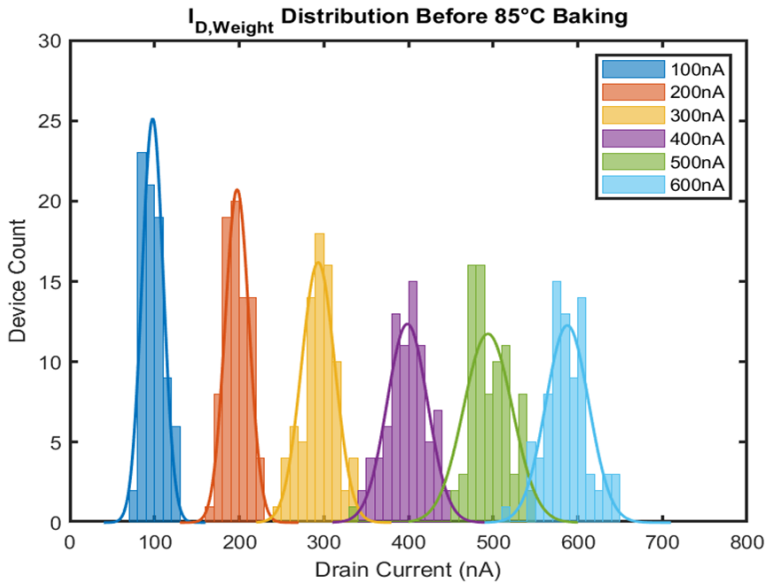


Fig. 3-5: A total of 480 devices are programmed to 6 different target levels (100,200,…,600nA). Each device is programmed to 3% within the target.

First of all, since all devices are programmed in array macros, neighboring cells are half-selected when a target cell is being programmed. Fig. 10 demonstrates this effect during PRG operation. When the target cell (circled in red) is programmed, all the cells in the same row (WL) see the same gate to source voltage bias (1.5-2.5V) which can cause partial PRG effect. All the cells in the same column also see the same drain to gate bias (2V) considering the unselected rows have 0V gate bias, which can cause partial ERS effect due to the reverse vertical field. Cells in an array experience these half-select effects numerous times before the entire array has been programmed. This is likely to be the major cause for the wide distribution of states after programming provided that the error of each cell is controlled to be less than or equal to 3%.

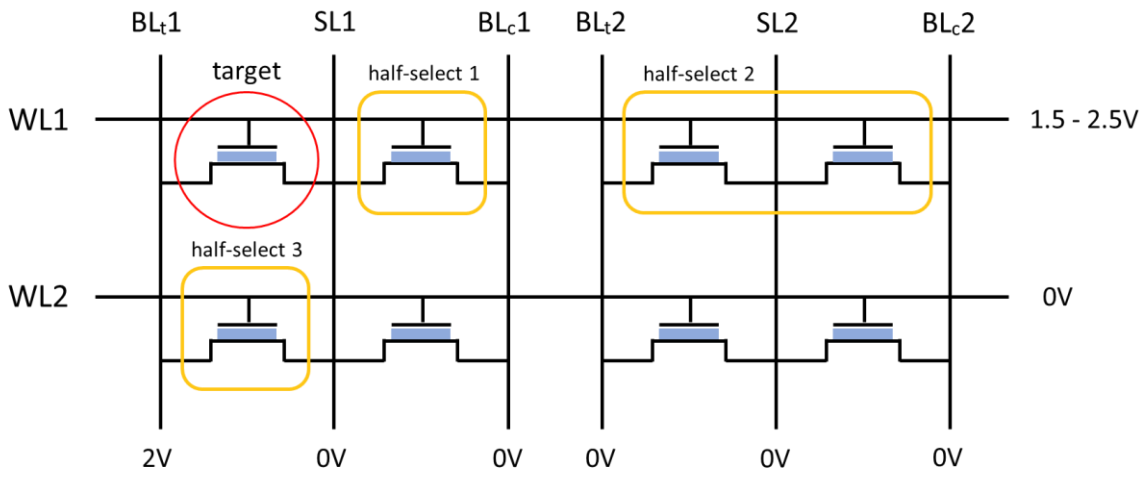


Fig. 3-6: Neighboring cell half-selection demonstration during PRG operation

The obtained distributions can be fitted to normal distribution curves and calculate the mean drift and standard deviation for each target level. Fig. 11 plots the data points of mean drift and standard deviation as a function of target level. Two quadratic curves are fitted to the data points. It is easy to see that both the mean drift and the standard deviation become worse as the target level goes up. However, the mean drift tends to deteriorate faster at higher target levels while the standard deviation tends to plateau.

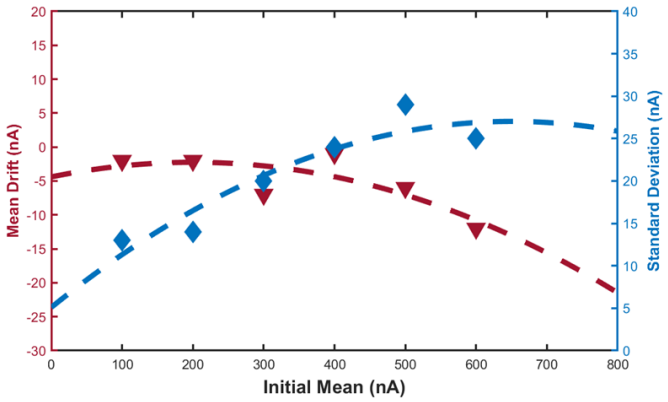


Fig. 3-7: Data points of mean drift and standard deviation of normal distribution curves for each target level. Quadratic curves are fitted to the data points.

Using this distribution model, the programming error for CTT can be approximated as a function of target cell current. The bit resolution, or equivalent number of bits (ENOBs), can consequently be modeled. Fig. 12 plots the estimated distribution given 2, 3, 4, and 5-bit resolution per device. It can be seen that devices can have distinct state distributions up to 3-bit/cell, and overlaps between states start to become noticeable at 4-bit. Resolution at 5-bit shows significant state overlapping.

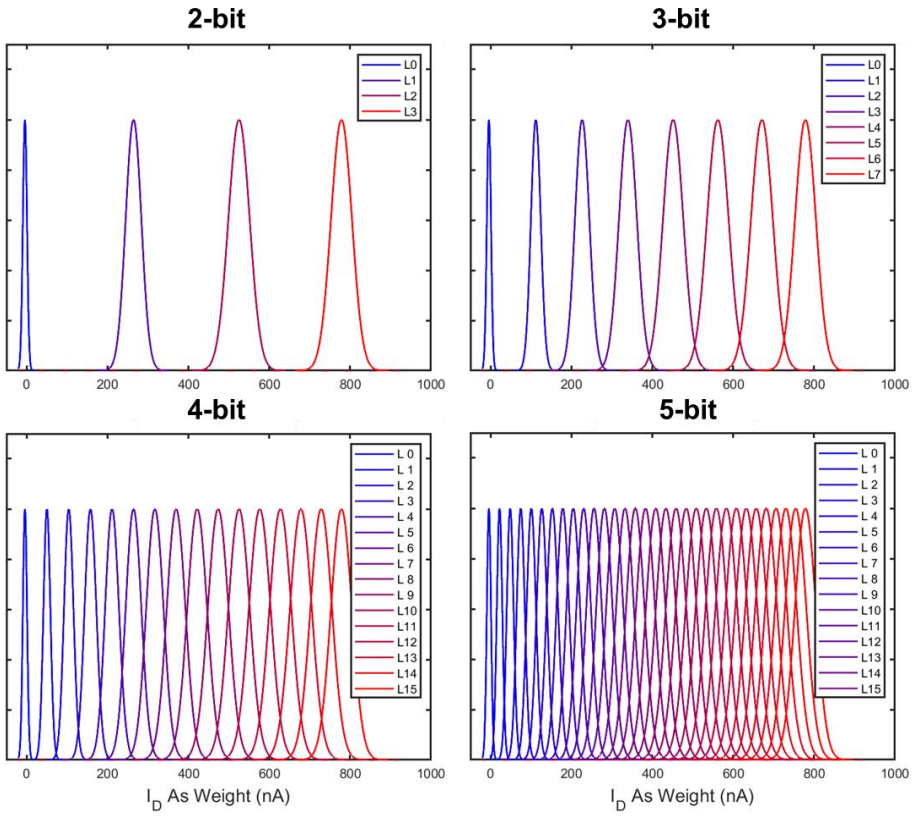


Fig. 3-8: Modeled ENOBs per CTT based on programming distributions obtained in fig. 3-5 and fig. 3-7. Fitted mean drift and standard deviations are used.

Chapter 4

Characterization of Device Retention

4.1 Overview

In general, retention characteristics of non-volatile memory devices are among the most significant device properties to be studied and evaluated. For digital applications, retention loss is measured as the data loss rate. For example, for Flash memory, the drift of the threshold voltage is monitored over time as a representation of data loss [9]. Although requirements for device retention may vary between different applications (automobile[10], biosensing[11]), in general more than 10 years retention at 85℃ is desired for digital applications. In the case of analog non-volatile memory, coming up with a standard can be more difficult because there is no representation of discrete levels in analog. However, similar characterization techniques can be utilized to observe the change in cell state over time.

This chapter starts by discussing CTT retention characteristics at room temperature. Then it proceeds to investigate the impact of different bias conditions used during programming on device retention at elevated ambient temperature, after which retention characteristics at 85℃ are studied. Finally, some limitations as well as the future perspectives of the work are discussed.

4.2 CTT Retention Characteristics at Room Temperature

In order to have a better understanding of CTT charge loss mechanism, drain currents of the devices under test have been continuously monitored for a long period of time. Fig. 13 (a) and (b) show examples of two devices being monitored for 20 hours. In both cases, currents are sampled every 10 seconds at 200mV gate and drain biases. The device in fig. 13 (a) has only been programmed with PRG pulses, so no "erase" is performed. Programming stops immediately once the current readout is lower than the target current. The device in fig. 13 (b) is programmed with both PRG and ERS pulses to fine-tune the current to be within 3% of the target level.

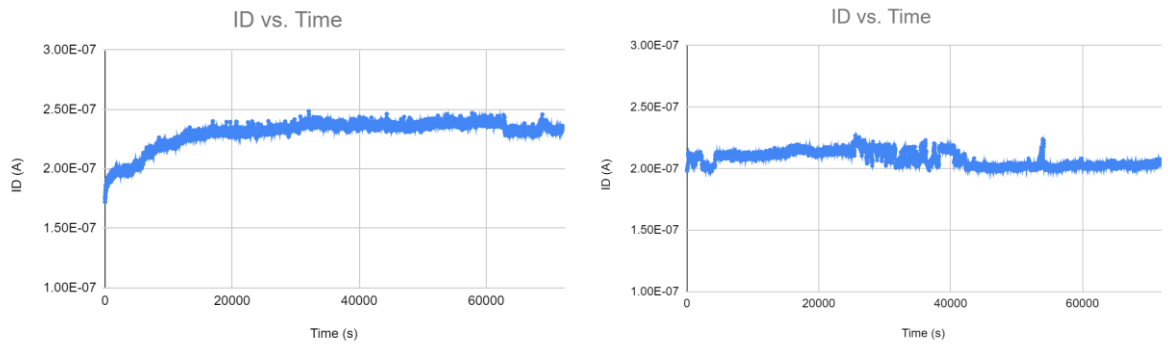


Fig. 4-1: (a) 20hr continuous current measurements of the CTT programmed with only PRG pulses; (b) 20hr continuous current measurements of the CTT programmed with both PRG and ERS pulses

A comparison between the two plots needs to be made to understand the two different retention phenomena more in depth. In case (a), the current rises quickly in the first few hours after programming, indicating that the charge trapped in low energy sites (shallow traps) are getting released, a phenomenon generally recognized as shallow trap relaxation. After the relaxation, the current rises slightly with intermittent "jumpy" behaviors. This can be explained

by the fact that more stable traps discharge at a slower rate compared to shallow traps with occasional re-trapping events. In case (b), the drain current exhibits a completely different trend than case (a). Instead of losing the trapped charge resulting in an increase in current, the device current remains generally flat across the entire 20-hour measurement period. However, noisy fluctuations occur frequently. The ERS pulses can help with getting rid of the charge in the shallow traps, which eliminates the initial rapid increase in current. Nevertheless, this process somehow has introduced noise into the device which becomes non-ideal for device operation in analog compute-in-memory applications.

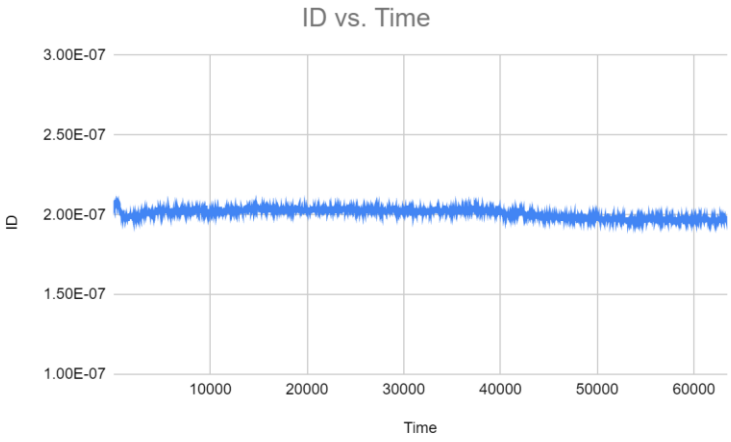


Fig. 4-2: 18hr continuous measurements on drain current after voltage burn-in

To reduce the noisy transients, a voltage burn-in method is introduced. After a device is programmed, 0.5V bias is applied to both the drain and the gate terminals to induce channel current for 20 minutes. Fig. 14 shows another device measured for 18 hours after voltage burn-in after programmed with both PRG and ERS pulses. Different from fig. 13 (b), the transients have been largely eliminated and the noise level in the current is significantly reduced. Thus, the voltage burn-in method can be added to be part of the device programming process. Using

the same programming/burn-in method as before, a total of six CTT devices are programmed to different target levels. Each device is then measured for 10 hours at room temperature. Their retention characteristics are shown in fig. 15, exhibiting excellent retention characteristics that can be extrapolated to the 10-year mark.

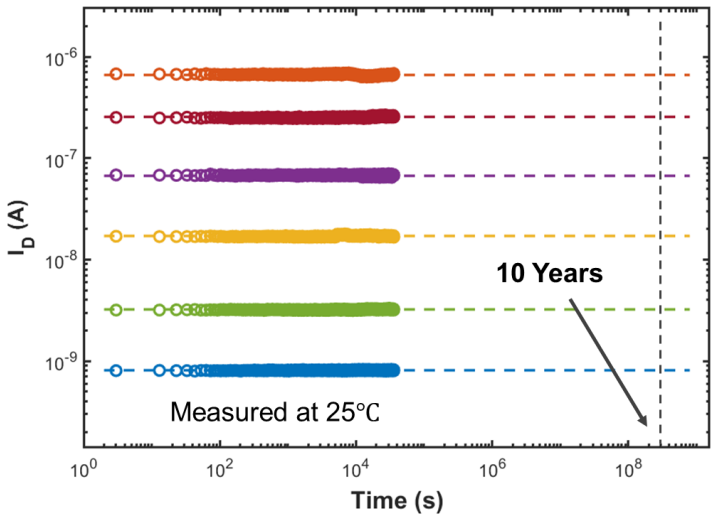


Fig. 4-3: Device retention at room temperature

4.3 Impact of Programming Voltage on CTT Retention at High Temperature

Due to the fact that the CTT in GlobalFoundries 22FDX has an ultra thin HfO_x/SiO_2 gate dielectric layer, it is reasonable to assume that the trapped charge can be somewhat leaky. Previous study on the trapping kinetics of the CTT shows that the self-heating process during programming helps with charge retention as trap sites with higher activation time constants can be activated with the elevated thermal energy [17 Faraz dissertation]. However, the correlation between programming voltage bias, specifically the drain-to-source bias, and retention characteristics at high temperature for analog applications, has not been studied before.

The design of experiments goes as follows. Three sets of devices are randomly selected with each set having six devices. All devices are programmed to the same target level, namely 200nA with the same gate voltage ramp scheme. Devices in set one, two, and three are programmed using 1.2V, 1.6V, and 2.0V drain-to-source biases, respectively. All devices are baked at 85℃ for 20 hours after programming.

Results of this experiment are summarized in table 7. After 20 hours at 85℃, it is obvious that all devices programmed by 1.2V drain-to-source bias suffer from significant charge loss with large variance among devices. The average percentage of the drift gets close to 100%, which is extremely non-ideal for analog computing applications. The mean drift percentage drops from 82.5 to 50.8 for devices programmed with 1.6V, which is still large considering the large change in a relatively short period of time. However, for all devices programmed with 2V lateral bias, the mean drift percentage is significantly reduced with much

less variance among devices. Fig. 16 plots the normalized mean drift and standard deviation of each group. It can be seen that at 2V bias, both mean drift and standard deviation are significantly reduced. Therefore, CTT retention characteristics at 85℃ exhibits a strong positive correlation with drain-to-source voltage bias used for programming. The higher the voltage, the more stable the trapped charge, the better the retention. However, the large drain bias in a transistor can accelerate device degradation, in that the increase in the lateral field can speed up hot carrier injection rates, making it more vulnerable to reliability issues.

@ VD=200mV, VG=200mV		Programming Target (nA)	After 20hr 85C (nA)	% Drift	Mean % Drift
VD = 1.2V	*Device 1*	200	367	83.5	82.5
	Device 2	200	430	115	
	Device 3	200	390	95	
	Device 4	200	323	61.5	
	Device 5	200	355	77.5	
	Device 6	200	325	62.5	
VD = 1.6V	*Device 1*	200	278	39	50.8
	Device 2	200	262	31	
	Device 3	200	325	62.5	
	Device 4	200	305	52.5	
	Device 5	200	297	48.5	
	Device 6	200	343	71.5	
VD = 2.0V	*Device 1*	200	207	3.5	-0.1
	Device 2	200	197	-1.5	
	Device 3	200	186	-7	
	Device 4	200	206	3	
	Device 5	200	194	-3	
	Device 6	200	209	4.5	

Table 4-1: Summary of the experiments with results

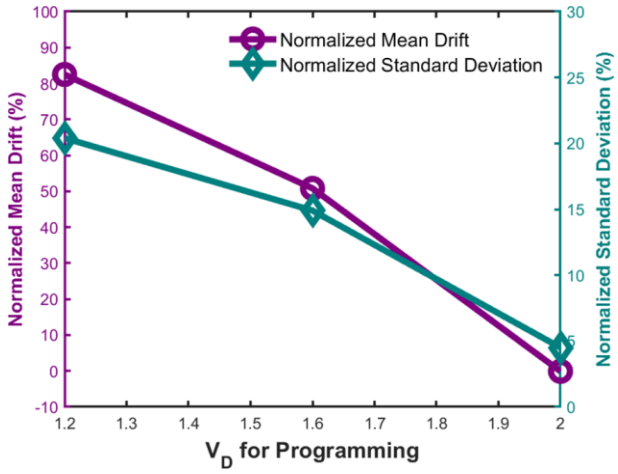


Fig. 4-4: Normalized group mean drifts and standard deviations with respect to the target level

4.4 CTT Retention Characteristics at High Temperature

The cell distributions in fig. 9 from Chapter 3 are obtained after performing voltage burn-in to all devices in the array. After this process is completed, devices are stored in an 85℃ environment and measured after 20 hours and 50 hours. Subsequent distributions are captured in fig. 17 (a), (b). For the purpose of comparison, the original distribution is given a transparent look overlaid by the actual distribution after baking. It can be noticed that although the cell distributions in both plots do not precisely overlap with each other, the change is rather insignificant. Fig. 18 summarizes the previous information and plots it in a whisker-box plot. This shows clear boundaries between the states.

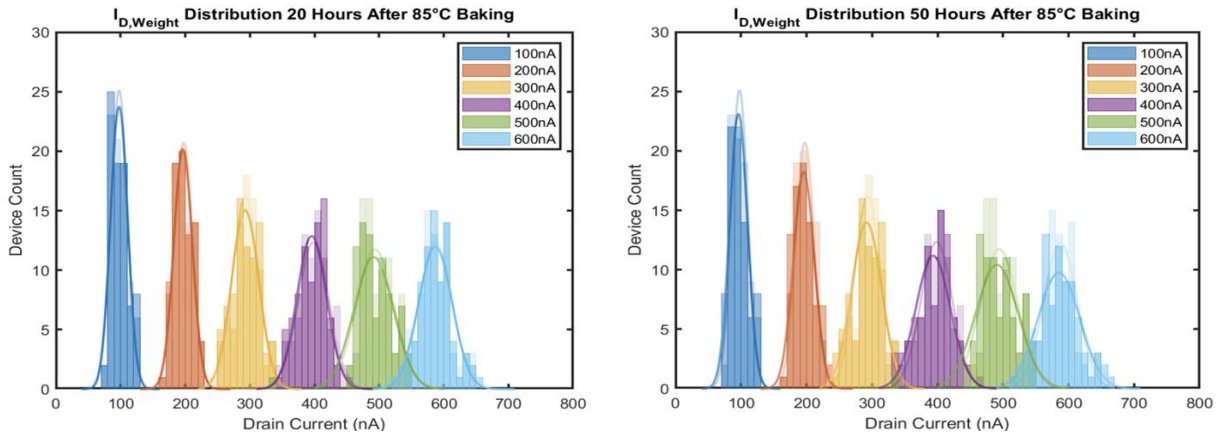


Fig. 4-5: (a) Cell distribution after 20 hours at 85℃; (b) Cell distribution after 50 hours at 85℃

Mean drifts and standard deviations of the distributions at T=20hr and T=50hr are also obtained from curve fitting. Fig. 19 provides data points and fitted curves of the two metrics at T=0, T=20hr, and T=50hr. In general, although the distributions are becoming wider with larger mean error as baking time increases, the change is rather insignificant. This is encouraging for applying the CTT to most types of analog in-memory compute schemes.

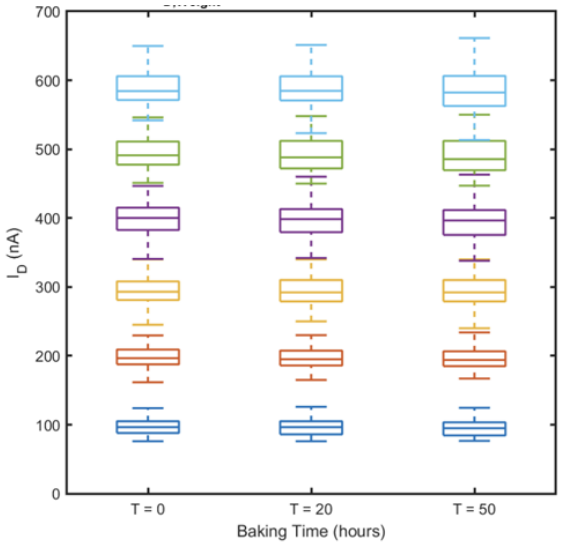


Fig. 4-6: Whisker-box plot for T=0, T=20hr, and T=50hr distributions

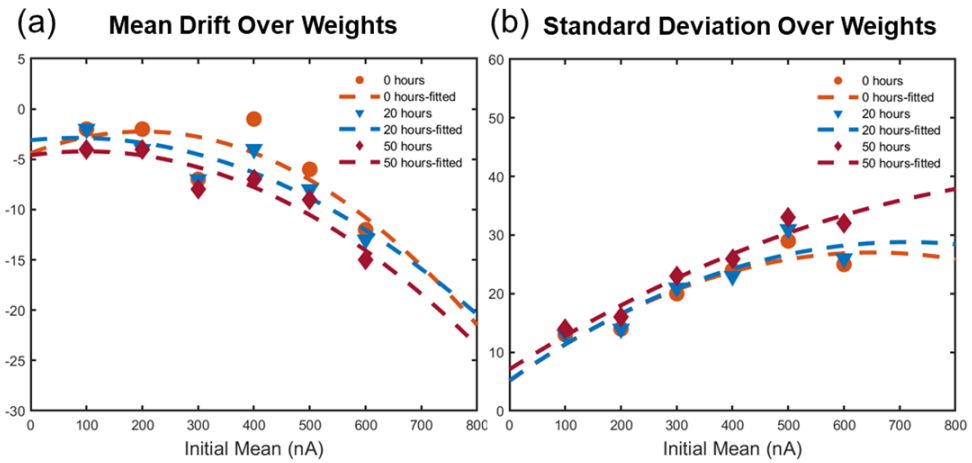


Fig. 4-7: (a) Measured and modeled mean drifts right after programming, after 20 hours at 85℃, and after 50 hours at 85℃. (b) Measured and modeled standard deviation right after programming, after 20 hours at 85℃, and after 50 hours at 85℃.

4.5 Limitations and Future Plans

Although the aforementioned results provide valuable insights into the CTT retention characteristics at the high temperature condition, they also come with a few limitations that need to be addressed carefully. First of all, all studies are done at 85℃. While this may be enough for some applications, it does not provide the full picture. Ideally, device retention should be characterized at different temperature points, ranging from -40℃ to 150℃. This would provide designers a much better view on what type of application CTT is suitable for. From a device perspective, it can also offer more insights into trapping and charge loss mechanisms. As an example, the CTT may not have as good retention properties at 125℃ as at 85℃, which is a desired metric for most automobile applications.

Second, baking time is limited to 50 hours, which is not enough for long-term retention characterization. Data points should be taken up until 1000 hours to help understand and model the cell distribution over time. The number of devices for testing can also be larger so that the statistics can be more convincing.

Future experiments should carefully address these issues. The design of experiments should find a balance between resources needed (time, number of devices, etc.) and the value of the potential results. It should also be made clear that the end goal of all experiments is two-fold: 1) to build statistically meaningful models, and 2) to provide physical insights.

Chapter 5

Summary

This work aims to provide a comprehensive device characterization of the CTT for analog compute-in-memory applications. Three aspects have been discussed in detail: noise-induced variation, programming accuracy, and retention characteristics.

In Chapter 2, it is suspected that device noise is related to the extent of programming and time of measurements. Experiments are designed such that variables are isolated from each other. From experiments, device noise is found to be consistent regardless of programming or when measurements are performed, showing little to no correlation to both variables.

In Chapter 3, the existing programming scheme and assumption and related results about error distribution are first laid out. A new programming scheme is then proposed to take advantage of both PRG and ERS pulses to be able to fine-tune the device level. The assumption is also challenged by the experiment which shows the distribution is a function of target states. A new distribution model is built based on the test results.

In Chapter 4, retention characteristics are first studied on discrete devices, finding the best way of programming to achieve better retention with less transients. Then, retention properties at room temperature and 85℃ become the focus of the study. Experiments suggest that the CTT has excellent retention at room temperature and show insignificant retention loss at 85℃. Limitations of the experiments are also discussed.