Title

Charge Trap Transistors (CTT): Turning Logic Transistors into Embedded Non-Volatile Memory for Advanced High-k/Metal Gate CMOS Technologies

Author

Khan, Faraz

Publication Date

2020

UNIVERSITY OF CALIFORNIA

Los Angeles

Charge Trap Transistors (CTT):

Turning Logic Transistors into Embedded Non-Volatile Memory for Advanced High-*k*/Metal Gate CMOS Technologies

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy

in Electrical and Computer Engineering

by

Faraz Khan

2020

ABSTRACT OF THE DISSERTATION

Charge Trap Transistors (CTT): Turning Logic Transistors into Embedded Non-Volatile Memory for Advanced High-*k*/Metal Gate CMOS Technologies

by

Faraz Khan

Doctor of Philosophy in Electrical and Computer Engineering

University of California, Los Angeles, 2020

Professor Jason C. S. Woo, Co-Chair

Professor Subramanian S. Iyer, Co-Chair

While need for embedded non-volatile memory (eNVM) in modern computing systems continues to grow rapidly, the options have been limited due to integration and scaling challenges as well as operational voltage incompatibilities. Introduced in this work is a unique multi-time programmable memory (MTPM) solution for advanced high-*k*/metal-gate (HKMG) CMOS technologies which turns as-fabricated standard logic transistors into eNVM elements, without the need for any process adders or additional masks. These logic transistors, when employed as eNVM elements, are dubbed "Charge Trap Transistors" (CTTs). The fundamental device physics, principles of operation, and technological breakthroughs required for employing

logic transistors as eNVM are presented. Implementation of CTT eNVM in 32 nm, 22 nm, 14 nm, and 7 nm production technologies has been realized and demonstrated in this work. The emerging memory technology landscape and the space that the CTT technology occupies therein are examined.

The motivation behind this work is to develop an eNVM technology that is completely process/mask-free, multi-time programmable, operable at low/logic-compatible voltages, scalable, and secure. The CTT technology satisfies all of the aforementioned criteria. CTTs offer a data retention lifetime of > 10 years at 125 °C and an operation temperature range of -55°-125° C. Hardware results demonstrate an endurance of > 10^4 P/E cycles which is more than adequate for most embedded applications. Hardware security enhancement, on-chip reconfigurable encryption, firmware, BIOS, chip ID, redundancy, repair at wafer and module test and in the field, performance tailoring, and chip configuration are a few of the applications of CTT eNVM. Moreover, the CTT array in its native (unprogrammed) state measures very well as an entropy source for potential PUF (Physically Unclonable Function) applications such as identification, authentication, anti-counterfeiting, secure boot, and cryptographic IP. In addition to the numerous digital applications, CTTs can also be utilized as an analog memory for applications like neuromorphic computing for machine learning (ML) and artificial intelligence (AI).

The dissertation of Faraz Khan is approved.

Chih-Kong Ken Yang

Yuanxun Wang

Jason C. S. Woo, Committee Co-Chair

Subramanian S. Iyer, Committee Co-Chair

University of California, Los Angeles

2020

DEDICATION

To my wife Florence and my daughter Eliza

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

ACKNOWLEDGEMENTS

I take this opportunity to express my sincere gratitude to my advisors Prof. Subramanian Iyer and Prof. Jason Woo for their invaluable guidance throughout my research work. Their constructive feedback, encouragement, and patience have been the driving forces for the successful completion of my research and dissertation.

I thank Prof. Ken Yang and Prof. Ethan Wang for providing a critical review of this work and for serving as members of my dissertation committee.

I am deeply grateful to Dr. DP Prakash (Global Head of Innovation, AI ML AR VR, CIO office at GLOBALFOUNDRIES) for his guidance and mentorship. I thank him for always taking time out for me despite his very busy schedule, encouraging me, and offering invaluable advice whenever I needed it the most.

I would like to express my sincere appreciation to Dr. Eduard Cartier (Research Staff Member at IBM T.J Watson Research Center) for his guidance and counsel. He always shared his vast technical knowledge, expertise, and experience to help me throughout my research work.

I am thankful to the management team at GLOBALFOUNDRIES, especially Mr. Norman Robson, Mr. Toshiaki Kirihata, and Mr. Daniel Berger, for supporting my research work and providing me with the necessary resources from conception to realization of the CTT eNVM technology.

I would like to thank DARPA and DMEA for their funding and support of this work.

I would like to especially thank my friend Dr. Gaurav Saraf for his encouragement and advice at every step of the way and for sharing his technical expertise throughout the course of my research work.

I express my deep sense of gratitude to my parents, my brother, and my sister for their never ending love and moral support. And finally, I thank my wife with all the appreciation in the world for always believing in me, for always being patient with me, and for never letting me give up.

VITA

PATENTS AND PUBLICATIONS

- F. Khan *et al*., “CHARGE TRAP MEMORY DEVICES”, U.S. Patent application # 16/781,527.
- F. Khan *et al*., “PROGRAM AND ERASE MEMORY STRUCTURES”, U.S. Patent application # 16/047,529.
- F. Khan, “Charge Trap Transistors (CTT): A Process/Mask-Free Secure Embedded Non-Volatile Memory for 14 nm FinFET Technologies and Beyond”, *Microelectronics Reliability and Qualification Workshop (MRQW)*, 2020 (Invited).
- F. Khan, D. Moy, D. Anand, E. Hunt-Schroeder, R. Katz, L. Jiang, E. Banghart, N. Robson, T. Kirihata, “Turning Logic Transistors into Secure, Multi-Time Programmable, Embedded Non-Volatile Memory Elements for 14 nm FINFET Technologies and Beyond”, *IEEE Symp. VLSI Technology*, Kyoto, Japan, 2019.
- F. Khan, M. Han, D. Moy, R. Katz, L. Jiang, E. Banghart, N. Robson, T. Kirihata, J. C. S. Woo, S. S. Iyer, "Design Optimization and Modeling of Charge Trap Transistors (CTTs) in 14 nm FinFET Technologies", *IEEE EDL*, 2019.
- F. Khan, E. Hunt-Schroeder, D. Moy, D. Anand, R. Katz, D. Leu, John Fifield, N. Robson, S. Ventrone, T. Kirihata, "A Multi-Time Programmable Embedded Memory Technology in a Native 14nm FINFET Process using Charge Trap Transistors (CTTs)," *GOMACTech*, 2019.
- F. Khan, E. Cartier, J. C. S. Woo, 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 EDL*, 2017.
- F. Khan, E. Cartier, C. Kothandaraman, J. Campbell, J. Woo, S. S. Iyer, “The Impact of Self-heating on Charge Trapping in High-*k*-metal-gate nFETs”, *IEEE EDL*, 2016.
- E. Hunt-Schroeder, D. Anand, D. Pontius, M. Roberge, D. Moy, E. Banghart, N. Robson, F. Khan, T. Kirihata, S. Ventrone, "Design and Development Challenges of Charge Trap Transistor Memories", *GOMACTech*, 2020.
- E. Hunt-Schroeder, D. Anand, J. Fifield, M. Roberge, D. Pontius, M. Jacunski, K. Batson, M. Deming, F. Khan, D. Moy, A. Cestero, R. Katz, Z. Chbili, E. Banghart, J. Liu, B. Jayaraman, R. Tummuru, R. Raghavan, A. Mishra, N. Robson, T. Kirihata, “14nm FinFET 1.5Mb Embedded High-K Charge Trap Transistor One Time Programmable Memory Using Differential Current Sensing”, *IEEE SSC-L*, 2019.
- B. Jayaraman , D. Leu, J. Viraraghavan, A. Cestero, M. Yin, J. Golz, R. Tummuru, R. Raghavan, D. Moy, T. Kempanna, F. Khan, T. Kirihata, S. S. Iyer, “80-kb Logic Embedded High-K Charge

Trap Transistor-Based Multi-Time-Programmable Memory With No Added Process Complexity", *IEEE JSSC*, 2018.

- 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", *IEEE Symp. VLSI Circuits*, 2016.
- C. Kothandaraman, X. Chen, D. Moy, D. Lea, S. Rosenblatt, F. Khan, D. Leu, T. Kirihata, D. Ioannou, G. LaRosa, J. B. Johnson, N. Robson, S. S. Iyer, "Oxygen vacancy traps in Hi-K/Metal gate technologies and their potential for embedded memory applications", *IEEE IRPS*, 2015.
- C. Kothandaraman, X. Chen, D. Moy, D. Lea, D. Leu, S. Rosenblatt, F. Khan, N. Robson, T. Kirihata, S. S. Iyer, "A secure CMOS compatible re-programmable memory element for code storage applications", *GOMACTech*, 2015.
- X. Yang, C. Ku, F. Khan, P. I. Reyes, C. Kuo, Y. Lu, "Sputtering of ZnO Thin Films for TFT on Polyimide Substrates", *Electronic Materials Conference*, 2010.
- C. Ku, P. I. Reyes, C. Kuo, F. Khan, "Annealing effects on ZnO TFTs grown by MOCVD", *6th International Workshop on ZnO and Related Materials*, 2010.

EMPLOYMENT

Technologist, 3D Memory R&D
Western Digital (SanDisk), Milpitas, CA
Dec 2019 – Present

Member of Technical Staff, Advanced Technology Development
GlobalFoundries, East Fishkill, NY
Jul 2015 – Dec 2019

Staff Scientist, Device Engineering + BEOL Technology Development
IBM, East Fishkill, NY
Nov 2011 – Jul 2015

Researcher
Microelectronic Research Lab (MERL), Rutgers University, Piscataway, NJ
Feb 2009 – Nov 2011

Lead Product Engineer, NAND Flash + DRAM
Micron Technology, Inc., Manassas, VA
Feb 2005 – Feb 2009

1. INTRODUCTION

1.1 MOTIVATION AND OBJECTIVES

The availability of on-chip non-volatile memory for advanced high-*k*/metal gate (HKMG) CMOS technology nodes has been constrained by integration and scaling challenges as well as operational voltage incompatibilities, while its need continues to grow rapidly in modern computing systems. Existing embedded memory solutions for HKMG CMOS technologies such as eFUSE [1], [2] and gate breakdown anti-fuse [3], [4] are one-time programmable and face scaling challenges. 1T1R eFUSE solutions, while operable at logic-compatible voltages, require a high current during the programming operation due to which the cell size must be several times larger than that of a logic transistor making them intangible for high density applications. eFUSEs therefore do not scale well in advanced and FinFET process technologies and when larger capacity one-time programmable memory (OTPM) is needed, the required area can be considerable. With the transition from traditional polysilicon-gate CMOS logic process to HKMG CMOS logic process in advanced technology nodes, the polysilicon eFUSE has been replaced by metal fuses, where the fuses are continuous metal shapes etched on the silicon surface. eFUSEs operate on the principle of electromigration (EM) and are programmed at the time of manufacturing: application of high voltages to selected fuse metal lines causes EM and subsequently disconnects (opens) the metal lines. However, eFUSEs suffer from high leakage current in the standby mode and from re-growth issues where the same electromigration that causes the metal lines to disconnect can also result in the metal lines to unintentionally connect again, changing the data intended to be stored. On the other hand, the anti-fuse technology employs an oxide-breakdown technique known to offer higher densities at the cost of using high voltages (>4V) which may be outside some technology limits or pose EMIR (electromigration/IR drop) concerns to periphery circuits due to the required high-density current flows. Additionally, reliably (irreversibly) breaking down ultra-thin gate oxides is becoming increasingly challenging, posing an additional constraint for anti-fuse technology scaling. Other solutions such as split-gate (SG) MONOS [5] and floating-gate (FG) type eFLASH [6], [7], [8] are multi-time programmable but require significant amount of additional masks and processing and require high voltages (up to ~10V) to operate. Moreover, scaling of FG eFLASH into FinFET technologies is unlikely due to significant process complexity and integration challenges. Emerging memory technologies such as MRAM [9], [10], [11], ReRAM [12], and PCM [13], [14], while multi-time programmable and generally operable at logic compatible voltages, require additional complex processes and masks.

The motivation of this work is to develop a multi-time programmable embedded non-volatile memory (eNVM) technology that is completely process-free/mask-free, operable at logic-compatible voltages (~2V), and scalable. The CTT technology, a novel multi-time programmable memory (MTPM) solution for advanced HKMG CMOS technology nodes which turns as-fabricated standard logic transistors into eNVM elements, satisfies all of the aforementioned criteria. While each of the technologies discussed earlier has its advantages and disadvantages, it must be noted that CTT is the only eNVM technology that is completely process- and mask-free i.e. it requires no additional processes or masks: this presents a significant time to market and cost advantage over all other memory technologies. Additionally, CTT eNVM offers a secure solution for data and hardware security. Data stored in eFUSE and anti-fuse memories can be reverse engineered (Fig. 1.1) using SEM voltage contrasting. On the other hand, data stored as trapped charge in a very thin dielectric (physical thickness to the order of ~ 1 nm or less) is very secure from reverse engineering and security attacks.

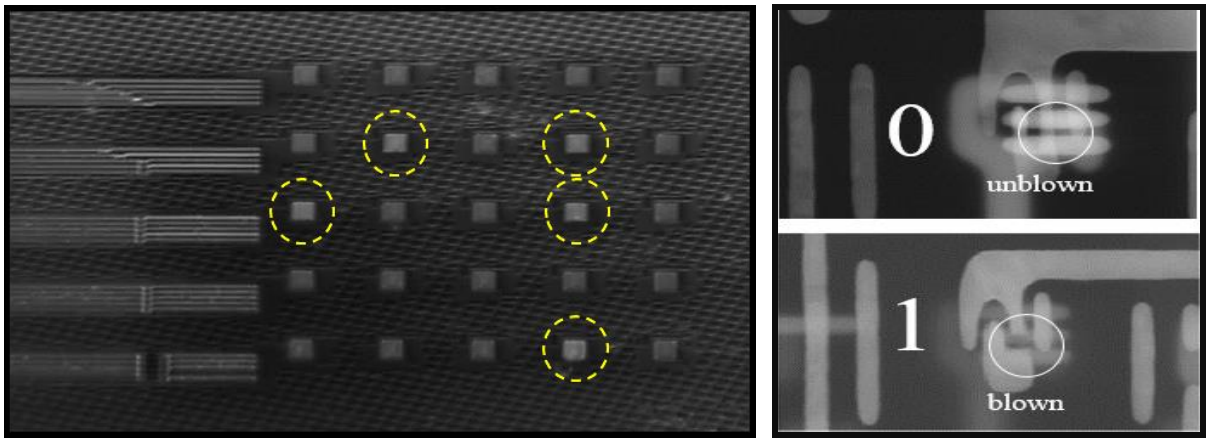
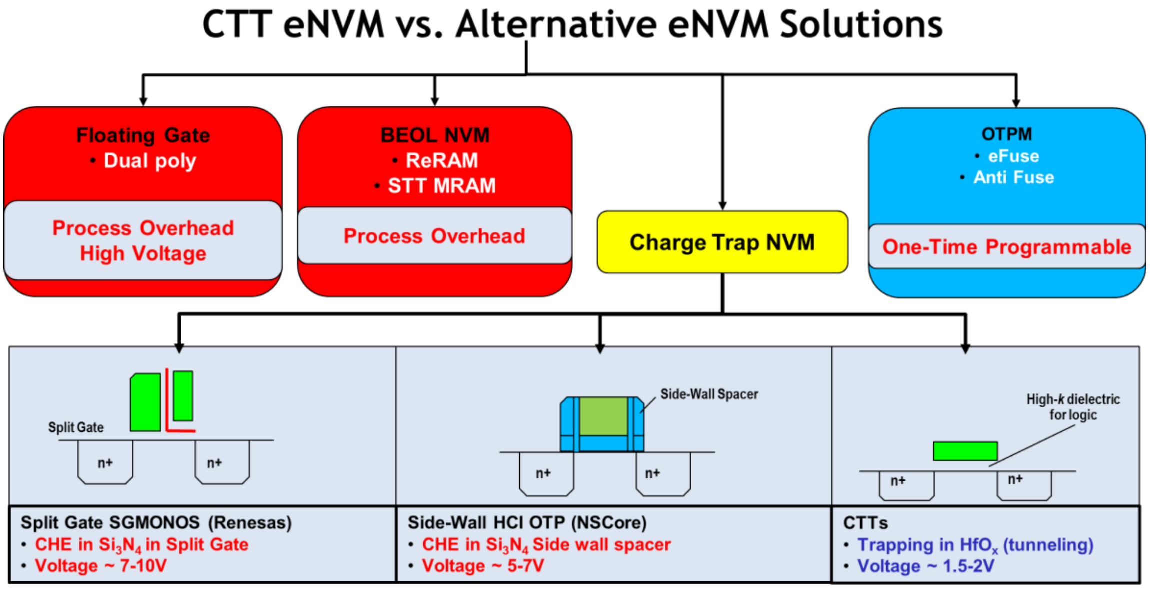


Fig. 1.1. SEM voltage contrasting showing blown vs. unblown anti-fuses (left) and a blown vs. unblown eFUSE (right), both in a 14 nm FinFET technology.

Potential applications of the CTT technology include hardware security, on-chip reconfigurable encryption, firmware, BIOS, chip ID, configuration memory, redundancy, repair at wafer and module test and in the field, and performance tailoring. Moreover, the CTT array in its native (unprogrammed) state measures very well as an entropy source for potential PUF (Physically Unclonable Function) applications such as identification, authentication, anti-counterfeiting, secure boot, and cryptographic IP. In addition to the numerous digital applications, CTTs can also be utilized as an analog memory for machine learning (ML) and artificial intelligence (AI) applications. A comparison between the various eNVM solutions and the CTT is shown in Fig. 1.2; The CTT is the only eNVM technology that is completely process-free / mask-free *and* multi-time programmable, operable at logic compatible voltages, secure, and scalable in bulk/SOI/FIN technologies. Given the eFLASH complexity and scaling challenges and no clear roadmap to sub-28 nm nodes, eFLASH replacement technologies are needed. A snapshot of the eNVM landscape and the emerging technologies for potential replacement of eFLASH for code/data storage and SRAM/eDRAM for working memory in 14 nm technology nodes and beyond, and where the CTT eNVM technology fits therein, is depicted in Fig. 1.3.

The objective of this work is to introduce the fundamental device physics and principles of operation of CTTs and to demonstrate viability of the CTT eNVM technology for advanced HKMG CMOS technology nodes. Also introduced are the technological breakthroughs required for employing CTTs as a commercially viable multi-time programmable eNVM technology, along with design and reliability considerations. Implementation of CTT eNVM in 32nm, 22nm, 14nm, and 7nm production technologies has been realized and demonstrated in this work.



CTT eNVM vs. Alternative eNVM Solutions

Floating Gate
- Dual poly

Process Overhead
High Voltage

BEOL NVM
- ReRAM
- STT MRAM

Process Overhead

Charge Trap NVM

OTPM
- eFuse
- Anti Fuse

One-Time Programmable

Split Gate SGMONOS (Renesas)
- CHE in Si_3N_4 in Split Gate
- Voltage ~ 7-10V

Side-Wall HCI OTP (NSCore)
- CHE in Si_3N_4 Side wall spacer
- Voltage ~ 5-7V

CTTs
- Trapping in HfO_x (tunneling)
- Voltage ~ 1.5-2V

Fig. 1.2. CTT eNVM vs. alternative eNVM solutions.

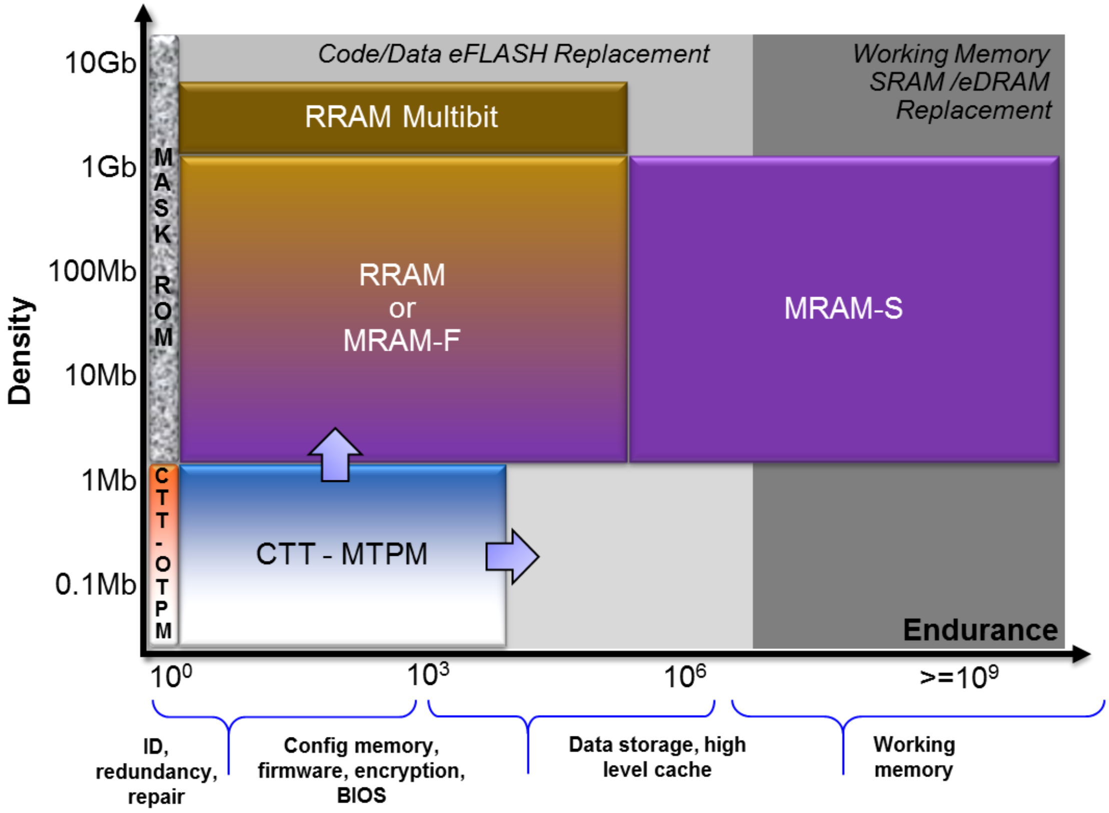


Fig. 1.3. Application mapping: Emerging eNVM landscape.

1.2 DISSERTATION ORGANIZATION

The motivation and objectives of this work are discussed in Chapter 1. Chapter 2 provides an introduction to and overview of the Charge Trap Transistor (CTT) eNVM technology for advanced HKMG CMOS technology nodes. A detailed discussion on the fundamental principles of operation of the CTT and its implementation as a one-time programmable memory (OTPM) is included in Chapter 3. The fundamental understanding and technological breakthroughs required for employing CTTs as a multi-time programmable memory (MTPM) are presented in Chapter 4. In Chapter 5, a compact model that can be used to accurately characterize and predict the behavior of CTTs and reliability considerations in the CTT eNVM technology are discussed. A summary of this work and corresponding conclusions and outlook are presented in Chapter 6.

REFERENCES

[1] C. Kothandaraman, S. K. Iyer and S. S. Iyer, “Electrically programmable fuse (eFUSE) using electromigration in silicides,” *IEEE Electron Device Lett.*, vol. 23, no. 9, pp. 523–525, Sep. 2002.

[2] S. H. Kulkarni, Z. Chen, B. Srinivasan, B. Pedersen, U. Bhattacharya and K. Zhang, “Low-Voltage Metal-Fuse Technology featuring a 1.6V-Programmable 1T1R Bit Cell with an Integrated 1V Charge Pump in 22nm Tri-gate process,” *IEEE Symp. VLSI Technol. Dig. Tech. Papers*, 2015, pp. C174-C175.

[3] S.-Y. Chou, Y.-S. Chen, J.-H. Chang, Y.-D. Chih and T.-Y. Chang, “A 10nm 32Kb Low-Voltage Logic-Compatible Anti-Fuse One-Time-Programmable Memory with Anti-Tampering Sensing Scheme,” *IEEE ISSCC*, 2017, pp. 200-202.

[4] Y. Liu, M. H. Chi, A. Mittal, G. Aluri, S. Uppal, P. Paliwoda, E. Banghart, K. Korablev, B. Liu, M. Nam, M. Eller and S. Samavedam, “Anti-Fuse Memory Array Embedded in 14nm FinFET CMOS with Novel Selector-Less Bit-Cell Featuring Self-Rectifying Characteristics,” *IEEE Symp. VLSI Technol. Dig. Tech. Papers*, 2014.

[5] S. Tsuda, T. Saito, H. Nagase, Y. Kawashima, A. Yoshitomi, S.Okanishi, T. Hayashi, T. Maruyama, M. Inoue, S. Muranaka, S. Kato, T. Hagiwara, H. Saito, T. Yamaguchi, M. Kadoshima, T. Maruyama, T. Mihara, H.Yanagita , K. Sonoda, T. Yamashita and Y. Yamaguchi, “Reliability and Scalability of FinFET Split-Gate MONOS Array with Tight Vth Distribution for 16/14nm-node Embedded Flash,” *IEEE IEDM*, 2017, pp. 19.3.1–19.3.4.

[6] R. Richter, M. Trentzsch, S. Dünkel, J. Müller, P. Moll, B. Bayha, K. Mothes, A. Henke1, M. Mazur, J. Paul, P. Krottenthaler, J. Poth, S. Jansen, R. Hüselitz, H. Kim, A. Zaka, T. Herrmann1, E.M. Bazizi, S. Beyer, P. Ghazavi, H. Om'mani, S. Lemke, Y. Tkachev, F. Zhou, J. Kim, X. Liu, V. Tiwari and N. Do, "A cost-efficient 28nm split-gate eFLASH memory featuring a HKMG hybrid bit cell and HV device," *IEEE IEDM*, 2018, pp. 428-431.

[7] Y. K. Lee, C. Jeon, H. Min, B. Seo, K. Kim, D. Kim, K. Min, J.S. Woo, H. Kang, Y.S. Chung, M. Kim, J. Jang, K.S. Yeom, J.-S. Kim, M.H. Oh, H. Lee, S. Cho and D. Lee, "High-Speed and Logic-Compatible Split-Gate Embedded Flash on 28-nm Low-Power HKMG Logic Process," *IEEE Symp. VLSI Technol. Dig. Tech. Papers*, 2017, pp. T202-T203.

[8] C. Jeon, J. Woo, K. Yeom, M. Seo, E. Hong, Y. Jeong, S. Lee, H. Min, D.H. Kim, H.C. Lee, S. Cho, M.H. Oh, J.-S. Kim, H. Lee, J.C. Park, C. Kim, H.J. Sung, S. Yoon, J. Kim, Y. K. Lee, K. C. Park, G. Jeong, J. Yoon and E.S. Jung, "High-speed and Ultra-low Power IoT One-chip (MCU + Connectivity-chip) on a Robust 28-nm Embedded Flash Process," *IEEE Symp. VLSI Technol. Dig. Tech. Papers*, 2019, pp. T114-T115.

[9] K. Lee, K. Yamane, S. Noh, V. B. Naik, H. Yang , S. H. Jang, J. Kwon, B. Behin-Aein, R. Chao, J. H. Lim, S. K., K. W. Gan, D. Zeng, N. Thiyagarajah, L. C. Goh, B. Liu, E. H. Toh, B. Jung, T. L. Wee, T. Ling, T. H. Chan, N. L. Chung, J. W. Ting, S. Lakshmipathi, J. S. Son, J. Hwang, L. Zhang, R. Low, R. Krishnan, T. Kitamura, Y. S. You, C. S. Seet, H. Cong, D. Shum, J. Wong, S. T. Woo, J. Lam, E. Quek, A. See and S. Y. Siah, "22-nm FD-SOI Embedded MRAM with Full Solder Reflow Compatibility and Enhanced Magnetic Immunity," *IEEE VLSI Technol. Dig. Tech. Papers*, 2018, pp. 183-184.

[10] K. Nishioka, H. Honjo, S. Ikeda, T. Watanabe, S. Miura, H. Inoue, T. Tanigawa, Y. Noguchi, M. Yasuhira, H. Sato and T. Endoh, “Novel Quad interface MTJ technology and its first demonstration with high thermal stability and switching efficiency for STT-MRAM beyond 2Xnm,” *IEEE Symp. VLSI Technol. Dig. Tech. Papers*, 2019, pp. T120-T121.

[11] O. Golonzka, J. -G. Alzate, U. Arslan, M. Bohr, P. Bai, J. Brockman, B. Buford, C. Connor, N. Das, B. Doyle, T. Ghani, F. Hamzaoglu, P. Heil, P. Hentges, R. Jahan, D. Kencke, B. Lin, M. Lu, M. Mainuddin, M. Meterelliyoz, P. Nguyen, D. Nikonov, K. O’brien, J. ODonnell, K. Oguz, D. Ouellette, J. Park, J. Pellegren, C. Puls, P. Quintero, T. Rahman, A. Romang, M. Sekhar, A. Selarka, M. Seth, A. J. Smith, A. K. Smith, L. Wei, C. Wiegand, Z. Zhang and K. Fischer, “MRAM as Embedded Non-Volatile Memory Solution for 22FFL FinFET Technology,” *IEEE IEDM*, 2018, pp. 412-415.

[12] O. Golonzka, U. Arslan, P. Bai, M. Bohr, O. Baykan, Y. Chang, A. Chaudhari, A. Chen, J. Clarke, C. Connor, N. Das, C. English, T. Ghani, F. Hamzaoglu, P. Hentges, P. Jain, C. Jezewski, I. Karpov, H. Kothari, R. Kotlyar, B. Lin, M. Metz, J. ODonnell, D. Ouellette, J. Park, A. Pirkle, P. Quintero, D. Seghete, M. Sekhar, A. Sen Gupta, M. Seth, N. Strutt, C. Wiegand, H. J. Yoo and K. Fischer, “Non-Volatile RRAM Embedded into 22FFL FinFET Technology,” *IEEE Symp. VLSI Technol. Dig. Tech. Papers*, 2019, pp. T230-T231.

[13] F. Arnaud *et al*., “Truly Innovative 28nm FDSOI Technology for Automotive Micro-Controller Applications embedding 16MB Phase Change Memory,” *IEEE IEDM*, 2018, pp. 424-427.

[14] P. Zuliani, A. Conte and P.Cappelletti, “The PCM way for embedded Non Volatile Memories applications,” *IEEE Symp. VLSI Technol. Dig. Tech. Papers*, 2019, pp. T192-T193.

2. Charge Trap Transistors (CTT): An Introduction and Overview

HfO_2 used as gate dielectric in high-*k*/metal gate (HKMG) CMOS technologies is known to have oxygen vacancy (Fig. 2.1) related charge traps [1], [2], [3]. An oxygen vacancy is a thermo-dynamic point defect caused by the diffusion of oxygen from HfO_2, which leaves behind a positively charged vacancy defect [4]. It is also known that bias stress induced charge trapping and defect generation in HfO_2 are strongly accelerated by temperature [5], [6]. While charge trapping in HfO_2 is typically considered to be a nuisance, as it is a source of variability in devices and in turn circuits, it is demonstrated in this work that this propensity for charge trapping in HfO_2 can indeed be utilized as a feature for embedded non-volatile memory (eNVM) applications in HKMG CMOS technologies. Charge trapping in high-*k* dielectrics such as HfO_2 for non-volatile memory (NVM) applications has been proposed before. In this work, however, it is shown that charge trapping in HfO_2 can be exploited for turning as-fabricated standard logic transistors into multi-time programmable (MTP) non-volatile memory elements that operate at logic compatible voltages without the need for any additional processes or masks: the fundamental device physics, principles of operation, and technological breakthroughs required are introduced. It is demonstrated that application of appropriate, logic compatible, voltages that are higher than nominal (~0.9V) can lead to enhanced charge trapping in the high-*k* gate dielectric material of HKMG logic transistors, resulting in threshold voltage shifts (ΔV_T) that are large and stable enough to be utilized as a non-volatile data storage mechanism. These as-fabricated standard logic transistors, when employed as eNVM elements, are dubbed “Charge Trap Transistors” (CTTs). In other words, CTTs are simply as-fabricated, standard HKMG CMOS logic transistors (Fig. 2.2) operated in an enhanced charge trapping mode.

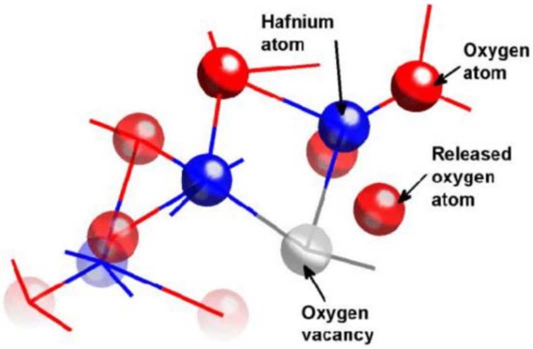


Fig. 2.1. Oxygen vacancy in HfO_2 generated by oxygen atom dislocation from the HfO_2 molecule [4].

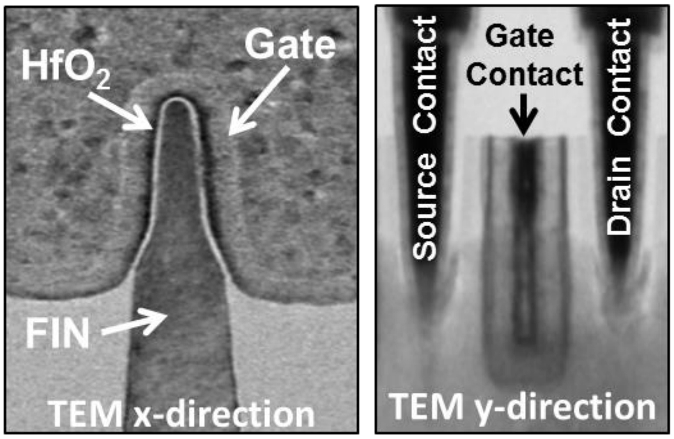


Fig. 2.2. TEM cross-sections of a 14nm FinFET CTT in the x- and y- directions (perpendicular and parallel to the FIN direction, respectively).

The fundamental principle of operation for programming CTTs is ‘device self-heating enhanced charge trapping’ [7], [8], [9]: the device threshold voltage (V_T) is modulated by charge trapped in the high-*k* dielectric of the HKMG device where the magnitude as well as stability (retention) of the trapped charge has a positive correlation to the self-heating temperature. The programming is typically done using short gate bias (V_G) pulses of ~1.8-2.0V with a drain bias (V_D) of ~1.3-1.6V, while the source bias (V_S) and substrate bias (V_X) are 0V (Fig. 2.3). The pulse of high channel current causes device self-heating while the high vertical field assists the electrons to inject into the gate and be trapped in the high-*k* dielectric, causing V_T to increase.

While detailed discussions are included in subsequent chapters, it is worthwhile pointing out here that the magnitude as well as the stability (retention) of the trapped charge is *significantly* enhanced due to the (self-heating induced) high temperature during the programming operation, resulting in large and stable V_T shifts suitable for NVM applications requiring high-temperature operation. The impact of device self-heating on the magnitude as well as the stability (retention) of the trapped charge is demonstrated in Fig. 2.4 and Fig. 2.5. Fig. 2.4 shows the measured ΔV_T from a CTT programmed, using the same programming voltage (V_G), with and without device self-heating and Fig. 2.5 shows the measured charge de-trapping activation energies (E_a) for devices programmed at various device self-heating temperatures: it is clear that the magnitude as well as charge retention characteristics are significantly enhanced by device self-heating during the charge trapping (programming) operation. The self-heating enhanced charge trapping is demonstrated to have excellent stability for the resulting device threshold voltage shifts to be used as a mechanism for non-volatile data storage: data retention lifetime of > 10 years at 125 °C has been demonstrated. Detailed discussions on maximizing the impact of device self-heating for improved operation efficiency and data retention for CTT memory are included in subsequent chapters.

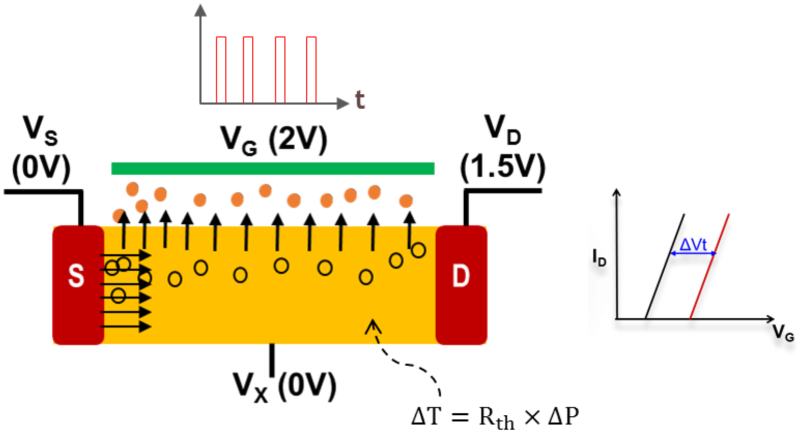


Fig. 2.3. A schematic of the CTT programming operation.

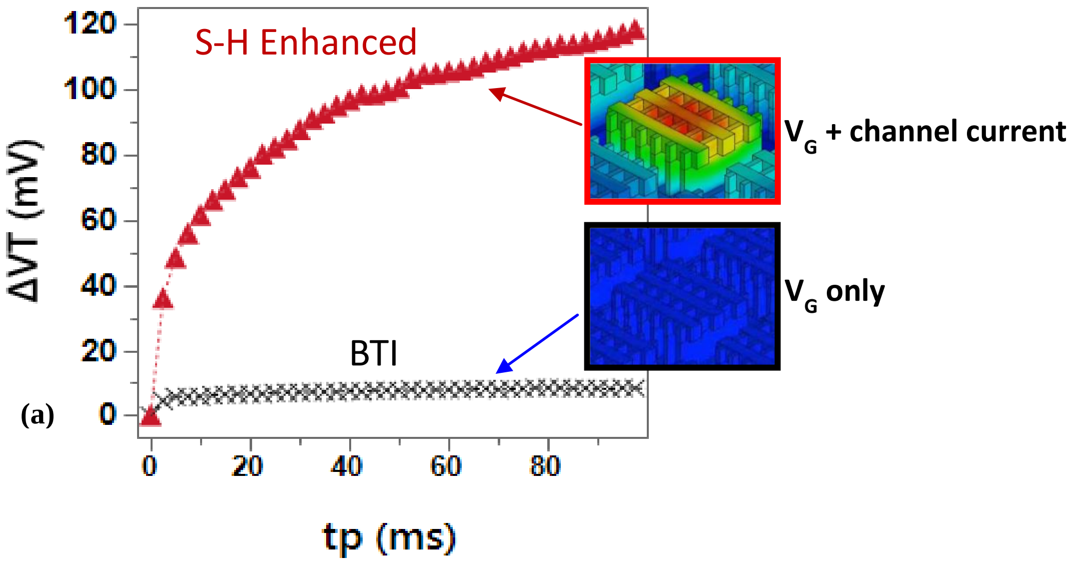


Fig. 2.4. Measured ΔV_T from a CTT programmed with and without device self-heating.

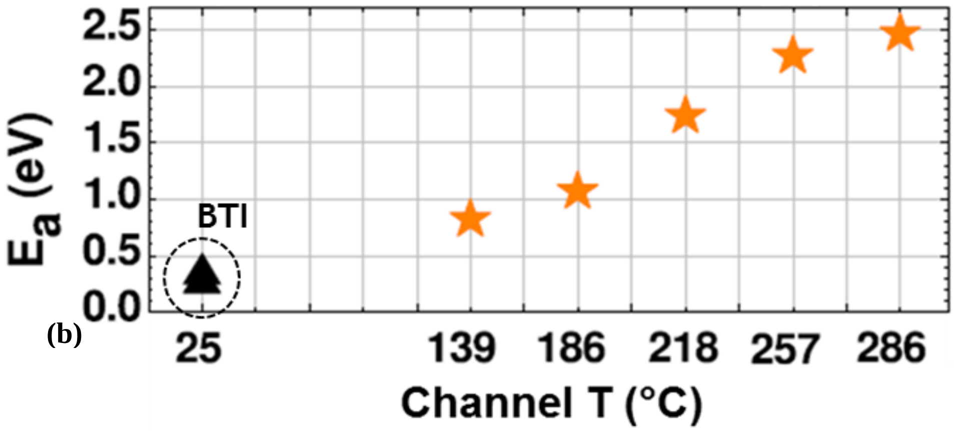


Fig. 2.5. Measured charge de-trapping activation energies (E_a) for CTTs programmed at various device self-heating temperatures.

The device V_T is modulated by the charge trapped in the high-*k* dielectric, whereafter each unique V_T value can be interpreted as a unique bit e.g. "0" and "1" for two unique V_T levels. The V_T of a transistor can be expressed by the following basic equation, where Q_{ox} is the quantity that is modulated due to the charge trapped in the gate dielectric:

$$V_{T0} = 2\Phi_F + \Phi_{ms} - \frac{Q_{ox}}{C_{ox}} - \frac{Q_{dm}}{C_{ox}} \tag{2.1}$$

where

Φ_F is the substrate Fermi potential (difference between the Fermi Level and the Intrinsic Fermi Level)

Φ_{ms} is the work function difference between the gate metal and the substrate

Q_{ox} is the trapped charge in the gate dielectric

Q_{dm} is the maximum charge held by the depletion layer at inversion

C_{ox} is the capacitance of the gate dielectric

It is clear that the effective threshold voltage of the device can be altered by the amount of charge in the high-*k* dielectric i.e. $V_T = V_{T0} + \Delta V_T$, where ΔV_T is given by $\Delta Q_{ox}/C_{ox}$.

The basic erase operation in CTT memory devices can be achieved by applying a negative gate-to-substrate bias, while the source, drain, and substrate are grounded, to electrostatically emit trapped charge. However, as discussed in subsequent chapters, this erase technique results in an inefficient erase: a technique called “Self-heating Temperature Assisted eRase” (STAR) has been developed [10], [11], [12] to address this issue and achieve high erase efficiency, which ultimately results in a *significant* improvement in the memory window and the program/erase cycling endurance of CTTs.

Fig. 2.6 shows a schematic depicting the basic operation of a CTT memory device. It must be noted that, while schematics of a planar device are shown for demonstration here, the phenomenon of intrinsic self-heating enhanced charge trapping is equally applicable to FinFET technologies. Self-heating in SOI planar technologies is of course higher as compared to bulk planar technologies. Self-heating in bulk FinFETs, while generally less than SOI FinFETs, is comparable to SOI planar devices and increases considerably with scaling [13], [14]: this makes the CTT technology highly scalable. Implementation of CTT eNVM has been demonstrated in 32nm SOI planar, 22nm SOI planar, 14nm SOI FinFET, 14nm bulk FinFET, and 7nm bulk FinFET technology nodes, including fully functional product prototype memory arrays. The CTT technology offers logic voltage compatible operation, scalability, high density (~0.144μm^2/bit for 22nm and ~0.082μm^2/bit for 14nm technology), and excellent retention (> 10 years @ 125 °C) for a fully integrated and scalable MTP eNVM that can be implemented without the need for any added process complexity or masks. In addition to being multi-time programmable, owed to the aforementioned advantages the CTT technology offers a better alternative to existing one-time programmable (OTP) technologies like eFUSE [15] and gate breakdown anti-fuse [16] as it can be used more effectively for yield improvement, chip configuration, redundancy, repair at wafer and module test and in the field , performance tailoring, and hardware security improvements such as chip ID and on-chip reconfigurable encryption key and firmware storage with lower power, higher density, and higher scalability, at no additional processing cost.

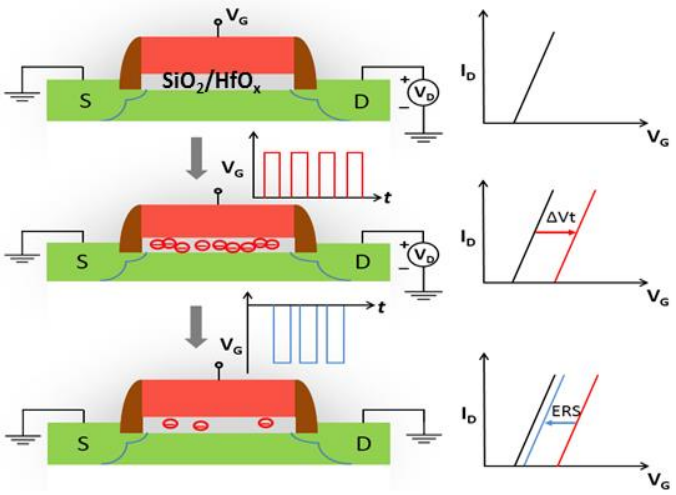


Fig. 2.6. A schematic depicting the basic operation of a CTT memory device (equally applicable to planar FET as well as FinFET based CTTs).

REFERENCES

[1] E. Cartier, B. P. Linder, V. Narayanan and V. K. Paruchuri, "Fundamental understanding and optimization of PBTI in nFETs with SiO_2/HfO_2 gate stack," *IEEE IEDM*, Dec. 2006, pp. 1-4.

[2] C. Kothandaraman, X. Chen, D. Moy, D. Lea, S. Rosenblatt, F. Khan, D. Leu, T. Kirihata, D. Ioannou, G. LaRosa, J. B. Johnson, N. Robson and S. S. Iyer, "Oxygen vacancy traps in Hi-K/metal gate technologies and their potential for embedded memory applications," *IEEE IRPS*, 2015, pp. MY.2.1–MY.2.4.

[3] H. Hamamura, T. Ishida, T. Mine, Y. Okuyama, D. Hisamoto, Y. Shimamoto, S. Kimura and K. Torii, "Electron trapping characteristics and scalability of HfO_2 as a trapping layer in SONOS-type flash memories," *IEEE IRPS*, 2008, pp. 412–416.

[4] A. R. Trivedi *et al.*, "A Simulation Study of Oxygen Vacancy-Induced Variability in HfO_2 /Metal Gated SOI FinFET," *IEEE Transactions on Electron Devices*, vol. 61, no. 5, pp. 1262-1269, May 2014.

[5] E. Cartier and A. Kerber, "Stress-induced leakage current and defect generation in nFETs with HfO_2/TiN gate stacks during positive-bias temperature stress," *IEEE IRPS*, 2009, pp. 486–492.

[6] F. Crupi, R. Degraeve, A. Kerber, D. H. Kwak and G. Groeseneken, "Correlation between stress-induced leakage current (SILC) and the HfO_2 bulk trap density in a SiO_2/HfO_2 stack," *IEEE IRPS*, 2004, pp. 181–187.

[7] F. Khan, E. Cartier, C. Kothandaraman, J. C. Scott, J. C. S. Woo and S. S. Iyer, “The Impact of Self-Heating on Charge Trapping in High- *k* -Metal-Gate nFETs,” *IEEE Electron Device Letters*, vol. 37, no. 1, pp. 88-91, 2016.

[8] 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, 2017.

[9] F. Khan, M. S. Han, D. Moy, R. Katz, L. Jiang, E. Banghart, N. Robson,T. Kirihata, J. C. S. Woo and S. S. Iyer, “Design Optimization and Modeling of Charge Trap Transistors (CTTs) in 14 nm FinFET Technologies,” *IEEE Electron Device Letters*, vol. 40, no. 7, pp. 1100-1103, 2019.

[10] F. Khan *et al*., “Program and Erase Memory Structures”, U.S. Patent app. # 16/047,529.

[11] F. Khan *et al*., “Charge Trap Memory Devices”, U.S. Patent app. # 16/781,527.

[12] F. Khan, D. Moy, D. Anand, E. H.-Schroeder, R. Katz, L. Jiang, E. Banghart, N. Robson and T. Kirihata, “Turning Logic Transistors into Secure, Multi-Time Programmable, Embedded Non-Volatile Memory Elements for 14 nm FINFET Technologies and Beyond,” *IEEE Symposium on VLSI Technology Digest of Technical Papers*, Kyoto, Japan, 2019, pp. T116-T117.

[13] T. Hook, F. Allibert; K. Balakrishnan, B. Doris, D. Guo; N. Mavilla, E. Nowak, G. Tsutsui, R. Southwick, J. Strane and X. Sun, “SOI FinFET versus bulk FinFET for 10nm and below,” *IEEE S3S Conf.*, 2014, pp. 1-3.

[14] D. Jang, E. Bury, R. Ritzenthaler, M. Garcia Bardon, T. Chiarella, K. Miyaguchi, P. Raghavan, A. Mocuta, G. Groeseneken, A. Mercha, D. Verkest and A. Thean, “Self-heating on bulk FinFET from 14nm down to 7nm node,” *IEEE IEDM*, 2015, pp. 6-11.

[15] C. Kothandaraman, S. K. Iyer and S. S. Iyer, “Electrically programmable fuse (eFUSE) using electromigration in silicides,” *IEEE Electron Device Lett.*, vol. 23, no. 9, pp. 523–525, Sep. 2002.

[16] Y. Liu, M.H. Chi, A. Mittal, G. Aluri, S. Uppal, P. Paliwoda, E. Banghart, K. Korablev, B. Liu, M. Nam, M. Eller and S. Samavedam, “Anti-Fuse Memory Array Embedded in 14nm FinFET CMOS with Novel Selector-Less Bit-Cell Featuring Self-Rectifying Characteristics,” *IEEE Symp. VLSI Technol. Dig. Tech. Papers*, 2014.

3. SELF-HEATING ENHANCED CHARGE TRAPPING AND CTT DESIGN OPTIMIZATION

In this chapter, the impact of device self-heating on the charge trapping behavior in high-*k*/metal-gate (HKMG) CMOS logic devices is studied, analyzed, and characterized. The magnitude of charge trapping is of course dependent on the applied gate bias i.e. the charge injection field. However, it is demonstrated that the channel temperature (T) during charge injection (programming), dictated by the device thermal resistance (R_{th}), also plays significant and perhaps a more important role in the charge trapping behavior. The phenomenon of self-heating enhanced charge trapping has been verified and studied on CTTs in several commercial technologies including 32 nm SOI planar, 22 nm SOI planar, 14 nm SOI FinFET, 14 nm bulk FinFET, and 7 nm bulk FinFET nodes, as demonstrated and discussed in subsequent chapters. CTTs in 22 nm SOI planar and 14 nm bulk FinFET nodes are used for demonstration purposes in this chapter.

The rise in device temperature during the CTT programming operation, or during any operation of any device for that matter, is given by the product of R_{th} and the applied power (P) i.e. $\Delta T = R_{th} \times P = R_{th} \times (I_{ch} \times V_D)$ where I_{ch} is the channel current and V_D is the applied drain-to-source bias. It is clear that the channel temperature can be increased by the applied power e.g. by increasing V_D and/or reducing the device channel length. However, in addition to an increase in self-heating temperature, a higher V_D and/or reduced channel length also results in an increased amount of hot carrier injection (HCI). In order to decouple the impact of the lateral electric field from the impact of temperature, device layout-dependent effects can be manipulated to strongly modulate and enhance the device R_{th} i.e. considerably higher device temperatures can be achieved for the same power and applied electric fields. In other words, it is demonstrated that the charge trapping is dependent not only on the channel power density during the programming operation, which is controlled by drain bias and device channel length, but it is also strongly modulated by the device layout. Thus, identical power densities in electrically identical devices (identical channel lengths and widths) with different device layouts, and different R_{th}, result in significantly different charge trapping behaviors. While device self-heating strongly influences the magnitude of charge trapping, it is found that the self-heating temperature during the charge injection (programming) operation also significantly enhances the stability (retention) of the trapped charge.

The implications of the findings for the application of high-*k*/metal-gate logic devices as embedded memory elements, dubbed as “Charge Trap Transistors” or “CTTs”, for non-volatile data storage in high-*k*/metal-gate CMOS technologies without added process complexity are discussed in this chapter. Considerations for optimization of bitcell design and operation conditions for CTT memory are also included.

3.1 Modulation of Self-Heating Enhanced Charge Trapping With Bias

For this study, experiments are performed on devices fabricated in a 22 nm high-performance SOI technology [1]. First, device threshold voltage shifts (ΔV_T) are measured during pulsed gate voltage ramp sweeps (PVRS) for various fixed drain bias (V_D) conditions. Gate bias (V_G) is applied using 10ms pulses of increasing magnitudes in 50mV increments. After each pulse, the device threshold voltage (V_T) is measured within 10ms. Each device is ramped until breakdown and Fig. 3.1 shows the measured ΔV_T values until before breakdown. Details on the PVRS technique can be found in [2]. The pre-stress V_T of each device is ~280mV. Two observations are made; Firstly, at higher V_D's (higher lateral field and self-heating), equivalent ΔV_T's are achievable at substantially lower V_G's. This is attributed to the impact of an enhanced level of HCI and charge trapping with increasing V_D as well as to enhanced charge trapping due to device self-heating [3] with increasing V_D. Secondly, the maximum achievable ΔV_T before device breakdown initially increases and then starts to decrease with increasing V_D. The breakdown of devices under low V_D conditions is electric field driven (high gate-to-drain bias, V_{GD}) whereas the breakdown of devices under high V_D conditions (which happens at much lower V_{GD}) is self-heating driven, which is a well-known phenomenon [4]. Shifts in ΔV_T vs. V_G trends before hard breakdown may be indicative of the beginning of soft breakdown [5].

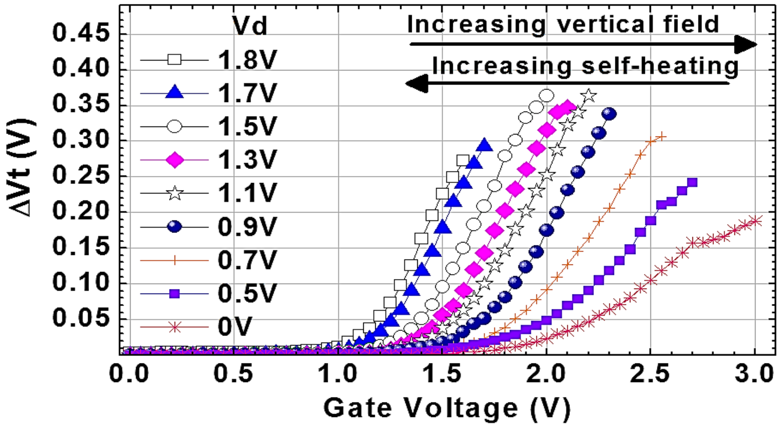


Fig. 3.1. ΔV_T as a function of PVRS stress with 10ms pulses at various fixed V_D values (W_{ch}=1.04 um, L=20 nm).

3.2 Exploiting Device Layout-Dependent Effects for Modulation of Self-Heating Enhanced Charge Trapping

While studying the impact of drain bias using the PVRS technique, as discussed in the previous section, is useful for understanding the dynamic charge trapping behavior as well as its bias dependence, that technique does not allow one to decouple the effect of the lateral electric field from the self-heating effects. In order to understand and quantify the impact of device self-heating on charge trapping in CTTs and to separate the impact of electric field from the thermal effects, layout-dependent effects are exploited to modulate the R_{th} of devices, while all other electrical parameters are kept constant.

To demonstrate the layout dependence of the thermal resistance and in turn self-heating enhanced charge trapping in planar devices, 22 nm SOI [1] CTTs with various geometries are studied. Identical program pulses (35ms at V_G=2V and V_D=1.3V) are applied to the same channel width (W_{ch}) but various channel lengths (L) are and the V_T's are measured within 10ms. It is seen that ΔV_T increases as L decreases (Fig. 3.2(a)), which is expected and consistent with increasing levels of hot carriers and self-heating (due to increase in lateral field) and decreasing V_T (due to short-channel effects) with decreasing L. However, when identical program pulses are applied to devices with the same L and various W_{ch}, it is observed that ΔV_T increases with W_{ch} (Fig. 3.2(b)). This phenomenon of ΔV_T varying with W_{ch} (while vertical and lateral fields and L are the same, and therefore the level of HCI is the same) is not readily explained by merely a field-dependent charge injection mechanism and is attributed to the impact of self-heating, which is strongly modulated by W_{ch}. Additionally, as demonstrated and discussed in detail later in this section, the negligible asymmetry between forward- and reverse-mode measurements after device programming at nominal conditions provides further evidence that drain-side HCI is indeed *not* the dominant charge injection mechanism in CTTs. The level of asymmetry, however, can be modulated by the applied V_D during the programming operation (high values of V_D result in a higher level of drain-side HCI). The reader is reminded that typically $V_G > V_D$ during the CTT programming operation. Throughout this work, the chuck temperature is always maintained at 25 °C unless otherwise stated.

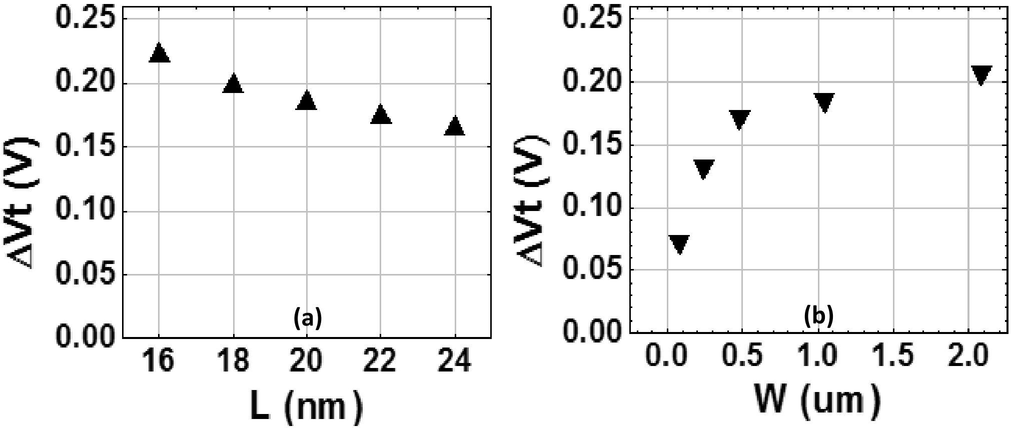


Fig. 3.2. Measured ΔV_T vs. (a) device channel length (W_{ch}=1.04 um) and (b) device channel width (L=20 nm).

To quantify the impact of layout to the device thermal resistance and demonstrate its subsequent impact on self-heating enhanced charge trapping, single-finger devices vs. multi-finger (split-channel) devices are studied. Both devices have a total W_{ch} of 1.04 um where each channel in the multi-finger devices, separated by trench isolation, has a width of $W_{ch}/4$. Both devices are identical to each other except for the channel width and have a channel length of 20 nm. Layouts of the two devices are shown in Fig. 3.3(a) and Fig. 3.3(b) respectively. First, channel thermal profiles of the two devices are analyzed. Thermal simulations have been carried out using finite element analysis (Comsol™). Full 3D structural simulations of the devices are analyzed and solved for temperature distribution and heat flux. Fig. 3.4(a) and Fig. 3.4(b) show the channel temperature (T) profiles of the two devices for an applied power density of 4mW/um. It is clear that W_{ch}/active area significantly modulates the device R_{th} and in turn self-heating. In multi-finger devices, the area for vertical heat flow is effectively larger than the area for power dissipation. Additionally, the larger area for lateral heat dissipation and the higher number of contacts per unit width in multi-finger devices are also responsible for higher heat dissipation and thus a lower R_{th} as compared to single-finger devices. The extracted R_{th} for the single-finger device is ~1.3x compared to the multi-finger device (65.9 vs. 50.8 K/mW, respectively). Additionally, simulation results show that the devices reach thermal equilibrium within ~200 ns (Fig. 3.5).

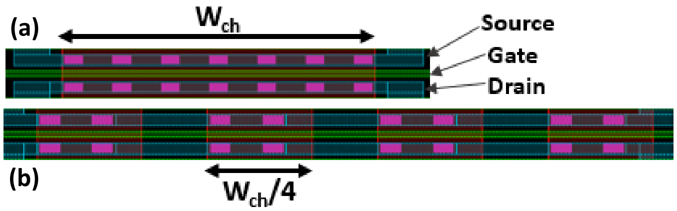


Fig. 3.3 (a) Single-finger and (b) Multi-finger device layout.

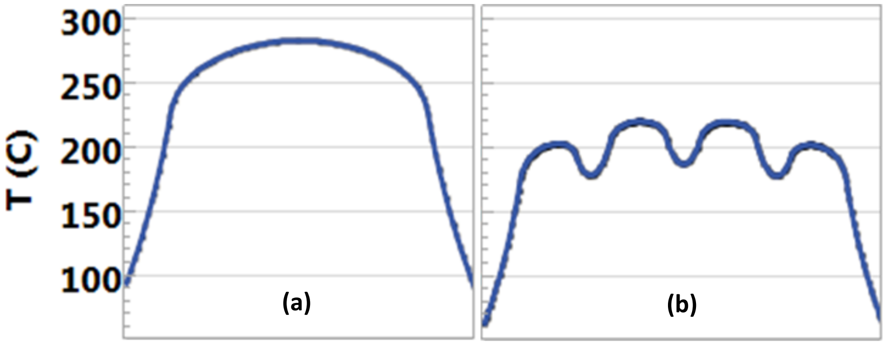


Fig. 3.4. Steady-state thermal profiles for a (a) single-channel and (b) multi-channel device in the W_{ch} direction, for an applied power of 4mW/um.

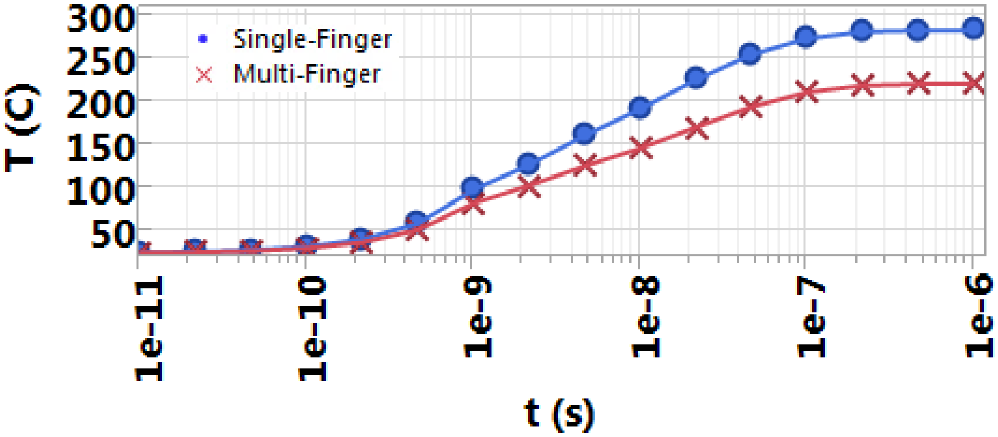


Fig. 3.5. Rise in channel temperature vs. time for self-heating (4mW/um applied power).

Fig. 3.6(a) shows the measured ΔV_T vs. applied power density for devices that were used for the thermal simulations. The power is varied by varying V_D while V_G=2V. It is seen that, for the same power density, ΔV_T for the single-channel device is considerably higher as compared to the split-channel device and the difference is greater at higher power densities. However, when plotted as a function of the calculated channel temperature (Fig. 3.6(b)), the ΔV_T characteristics of the two devices are almost identical except at very high temperatures where the single-channel device seems to have slightly higher ΔV_T. In other words, ΔV_T behaviors of the devices show a very strong correlation to the self-heating temperature as opposed to the applied power density. It is clear from these results that the device self-heating temperature is a significant factor in modulating the charge trapping behavior.

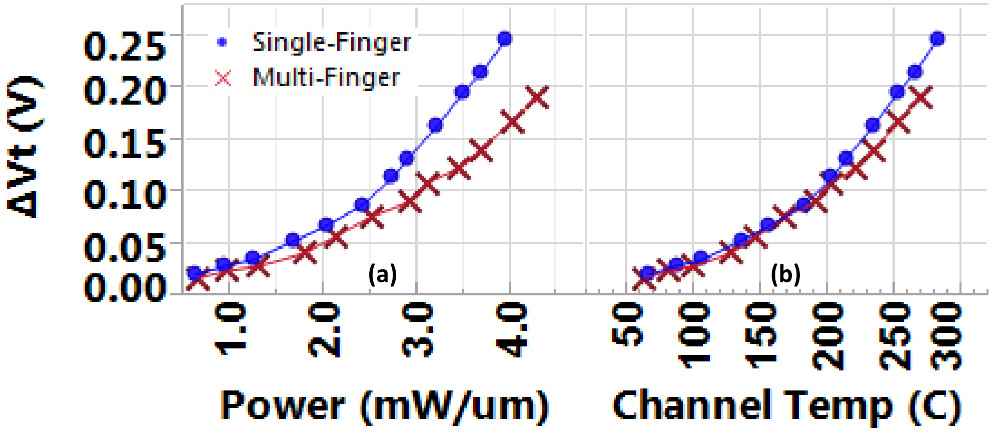


Fig. 3.6. Measured ΔV_T vs. (a) applied power density and (b) channel T during programming. It is observed that, at higher T's, the rate of increase in ΔV_T is higher.

Thus far, it has been shown how layout-dependent effects in planar devices can be manipulated to modulate and enhance the self-heating effect and in turn the programming efficiency in CTTs - Device self-heating (or alternatively thermal resistance, R_{th}) and therefore programming efficiency is strongly influenced by the width of each active channel in the planar devices: a single wide channel device shows a considerably higher programming efficiency as compared to a device with multiple narrower channels in parallel.

Next, how the CTT bitcell design can be manipulated to exploit layout-dependent effects significantly enhance the programming efficiency in FinFET-based CTTs is demonstrated; experiments are performed on hardware in a 14 nm FinFET technology [6]. Nominal nFET devices with a gate length of 14 nm and EOT of ~1.3 nm are used.

Unlike planar devices, the width of each active channel in FinFET devices is quantized i.e. the channel width of a device can only be increased by connecting multiple fins, and therefore a single channel cannot be made wider to increase the device R_{th}. However, the efficiency of thermal dissipation, and in turn the R_{th}, of FinFET devices can be modulated by changing the aspect ratio of the device i.e. by reconfiguration of the number of fins-to-number of gates ratio in each device**.** Another way to modulate the device R_{th} is by isolating bitcells from each other.

In order to optimize the bitcell layout to improve the effect of device self-heating and in turn the programming efficiency of CTTs in FinFET technologies, four different bitcell layouts are fabricated and studied. A ‘1 gate × 12 fin’ bitcell (Fig. 3.7(a)) and a ‘2 gate × 6 fin’ bitcell (Fig. 3.7(b)) are investigated. In addition to the bitcell aspect ratio, we investigate the impact of isolating the bitcells from each other, i.e. each bitcell is fabricated on an active “island” separated by a trench isolation. Fig. 3.7 (c) and Fig. 3.7 (d) show the isolated ‘1 gate × 12 fin’ and ‘2 gate × 6 fin’ bitcells, respectively. It must be noted that the only difference between each bitcell is the layout: each bitcell is composed of 12 FETs. The bitcells are programmed at V_G = 2V, V_D = 1.4V, and V_S = 0V using 2.5 ms pulses and the V_T’s are measured after each pulse. In order to study the charge trapping behavior in the absence of self-heating (no channel current, I_{ch}), devices are also programmed at $V_G = 2V$ and $V_D = V_S = 0V$. In order to examine the thermal dissipation properties of the corresponding bitcell designs, 3D finite element thermal simulations, using Sentaurus Interconnect, are also performed. For each bitcell, a power density of 7.1×10^{12} W/cm^3 associated with the Joule heating produced from current flow in the active fin channels during programming is applied and the respective channel temperatures and R_{th} values are extracted.

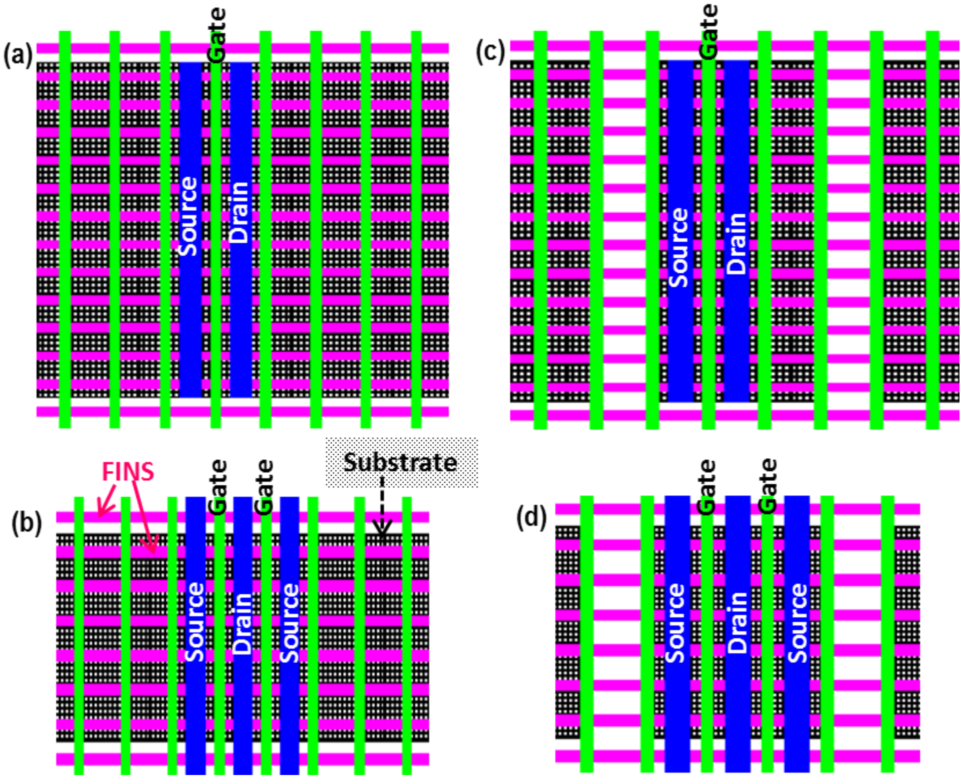


Fig. 3.7. Top down views of (a) ‘1 gate × 12 fin’, (b) ‘2 gate × 6 fin’, (c) ‘1 gate × 12 fin’ on active “island”, and (d) ‘2 gate × 6 fin’ on active “island” CTT bitcell layouts.

The bitcell steady-state temperatures, achieved within ~50 ns (Fig. 3.8), during the programming operation are shown in Fig. 3.9. The thermal profiles along the gate direction (perpendicular to the fins), at programming conditions, of each of the four bitcells (Fig. 3.10) show that the ‘2×6’ layout has a higher R_{th} and hence, for identical power densities, a higher channel temperature as compared to the ‘1×12’ layout. Furthermore, isolated bitcells have a higher R_{th} as compared to their un-isolated counterparts. Measured (hardware) data for the V_T shift (ΔV_T) vs. programming time (t_P) for each of the fabricated bitcell designs is shown in Fig. 3.11. From the measured hardware data and the corresponding thermal simulations, we make two key observations: First, in the presence of self-heating, bitcells with different layouts (and in turn R_{th}) exhibit considerably different behaviors with identical programing conditions. With the isolated ‘2×6’ bitcell, ΔV_T for the same t_P increases > 60%, > 30%, and > 10% as compared to the unisolated ‘1×12’ bitcell, the isolated ‘1×12’ bitcell, and the unisolated ‘2×6’ bitcell, respectively. The isolated ‘2×6’ bitcell enables a 6× reduction in t_P to reach the target ΔV_T as compared to the unisolated ‘1×12’ bitcell (Fig. 3.11 (a)). Secondly, in the absence of self-heating (Fig. 3.11 (b)), ΔV_T is negligible for the same V_G and that all the devices behave identically. These results are consistent with the findings in 32 and 22 nm planar SOI CTTs.

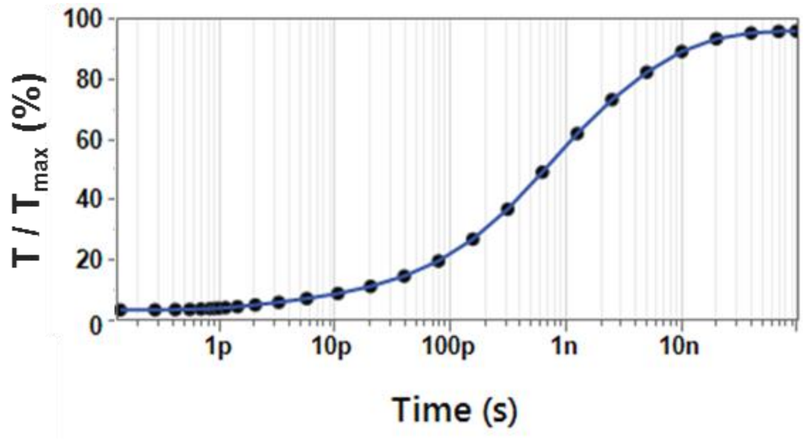


Fig. 3.8. Rise in bitcell temperature vs. time during the program operation.

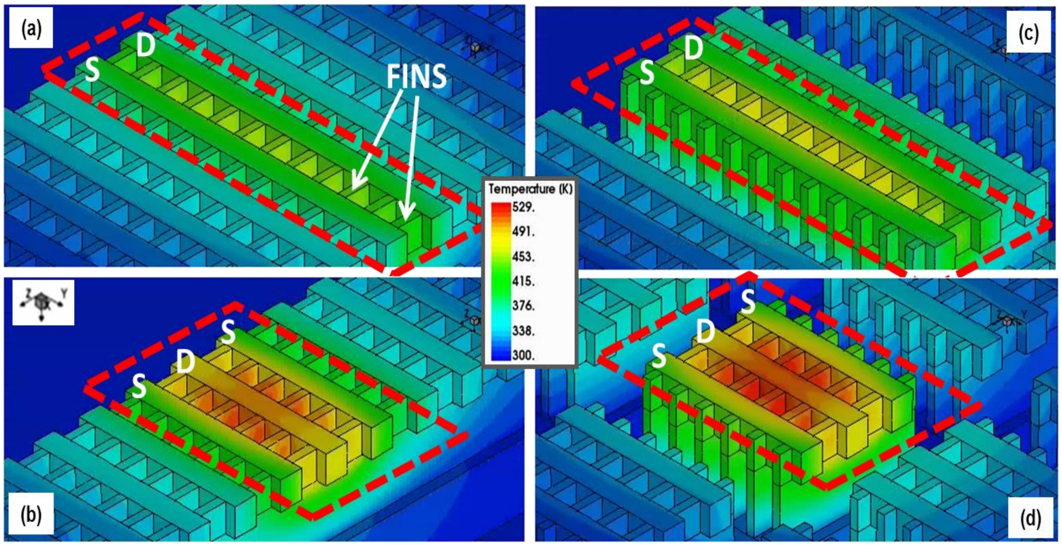


Fig. 3.9. 3D finite element thermal simulation of the programming operation of CTT bitcell structures with (a) ‘1 gate×12 fin’, (b) ‘2 gate×6 fin’, (c) ‘1 gate×12 fin’ on active “island”, and (d) ‘2 gate×6 fin’ on active “island” layouts.

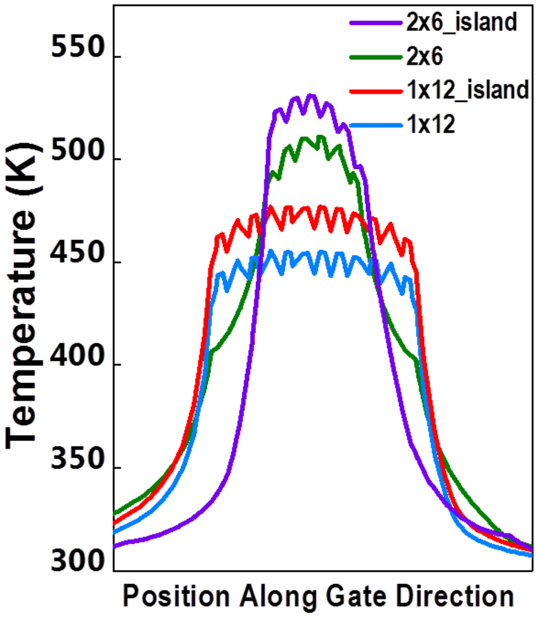


Fig. 3.10. Thermal profiles, during programming, of the bitcell layouts shown in Fig. 3.9, along the gate direction.

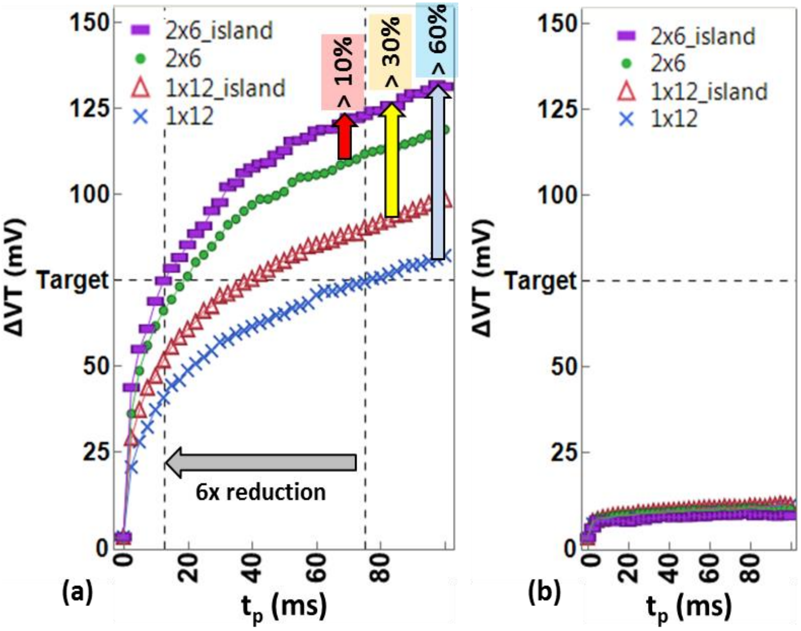


Fig. 3.11. ΔV_T vs. t_P for CTT bitcells with various layouts. The devices are programmed with (a) V_G=2V, V_D=1.4V, V_S=0V and (b) V_G=2V, V_D=V_S=0V.

3.3 Charge Injection Mechanism and the Charge Trapping Profile

In order to understand the charge injection behavior and mechanism(s), charge injection currents during the programming operation are measured at the gate terminal of 14 nm FinFET CTTs. Fig. 3.12(a) shows the increase in normalized charge injection currents measured, during programming using various power densities, from devices with different R_{th}. Similar to what is observed for the ΔV_T trends, for the same power density the charge injection current in the single-channel device (higher R_{th}) is considerably higher as compared to the split-channel device (lower R_{th}). However, when viewed as a function of the calculated channel temperature (Fig. 3.12(b)), the charge injection current characteristics of the two devices are very similar. In other words, the magnitude of charge injection currents shows a very strong correlation to the device self-heating temperature as compared to the applied power density. This observation reaffirms the conclusion that device self-heating temperature during the programming operation is a significant factor in modulating the charge injection and, in turn, the charge trapping behavior of CTTs.

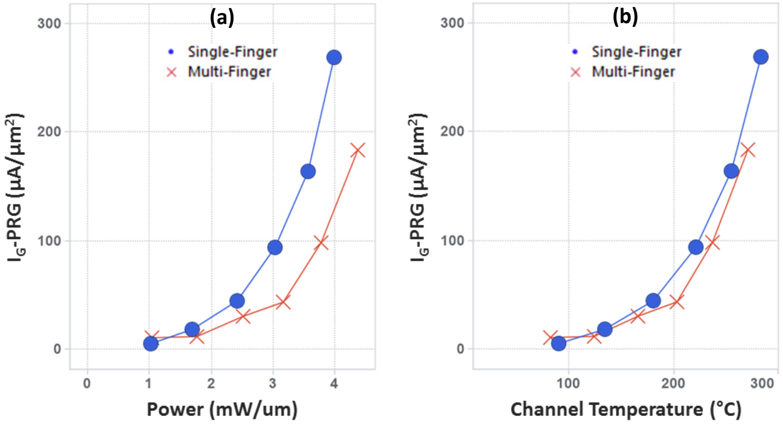
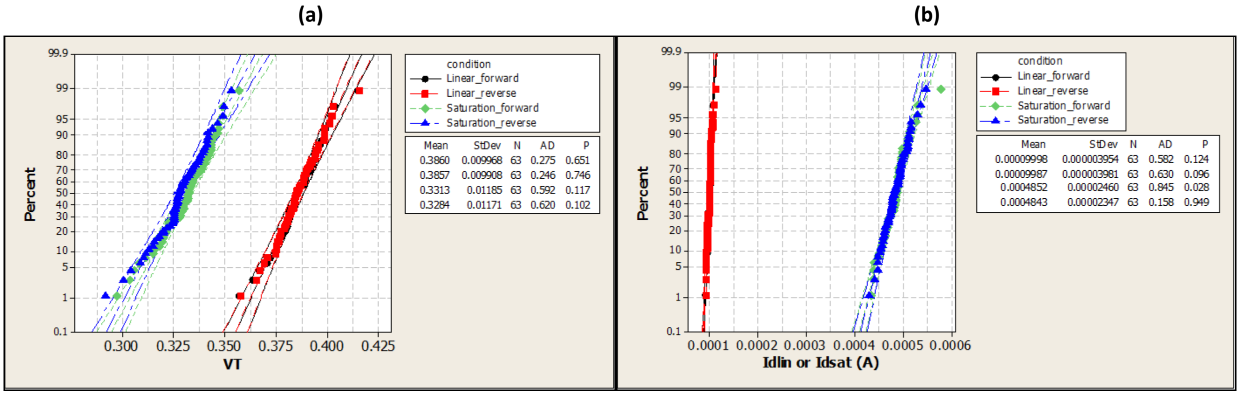


Fig. 3.12. Increase in charge injection current during programming vs. (a) applied power density and (b) channel T during the programming operation.

Direct tunneling through thin dielectrics is a strong function of temperature. However, given that the HfO_2 in the gate dielectric has oxygen vacancies that serve as charge trapping centers, the large temperature dependence of the gate current also indicates the presence a Poole-Frenkel (PF) conduction mechanism [7]. Additionally, Schottky–Richardson (SR) emission [8], which is a thermionic emission of an electron jump over a surface barrier, might also possibly be present during the CTT programming operation.

It is found that there is no significant polarity effect on the mean value of device V_T's and saturation currents (Fig. 3.13, Fig. 3.14), suggesting that, while some asymmetry might be present, overall the trapped charge is fairly uniformly distributed along the channel. The corresponding stochastic variation results in a small standard deviation (2.8%) for normalized deltas between forward- and reverse-mode reads (Fig. 3.14). Forward- and reverse-mode V_T

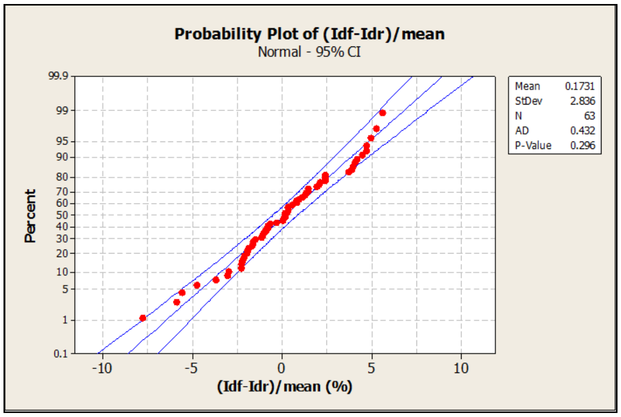
measured during a programming operation (Fig. 3.15) and post-program forward- vs. reverse-mode transconductance (Gm) measurements (Fig. 3.16) also show very little asymmetry, again indicating that the trapped charge distribution along the channel is fairly uniform. These results, once again, suggest that drain-side HCI is not the dominant charge injection mechanism in CTTs. Similar results have been observed for 22 nm and 32 nm SOI CTTs.



Mean	StDev	N	AD	P
0.3860	0.009968	63	0.275	0.651
0.3857	0.009908	63	0.246	0.746
0.3313	0.01185	63	0.592	0.117
0.3284	0.01171	63	0.620	0.102

Mean	StDev	N	AD	P
0.00009998	0.000003954	63	0.582	0.124
0.00009987	0.000003981	63	0.630	0.096
0.0004852	0.00002460	63	0.845	0.028
0.0004843	0.00002347	63	0.158	0.949

Fig. 3.13. Reverse- vs. forward-mode distributions for linear and saturation (a) V_T and (b) channel currents for 14 nm FinFET CTTs.



Mean	0.1731
StDev	2.836
N	63
AD	0.432
P-Value	0.296

Fig. 3.14. Stochastic variation in standard deviation for normalized deltas between forward- and reverse-mode reads for 14 nm FinFET CTTs.

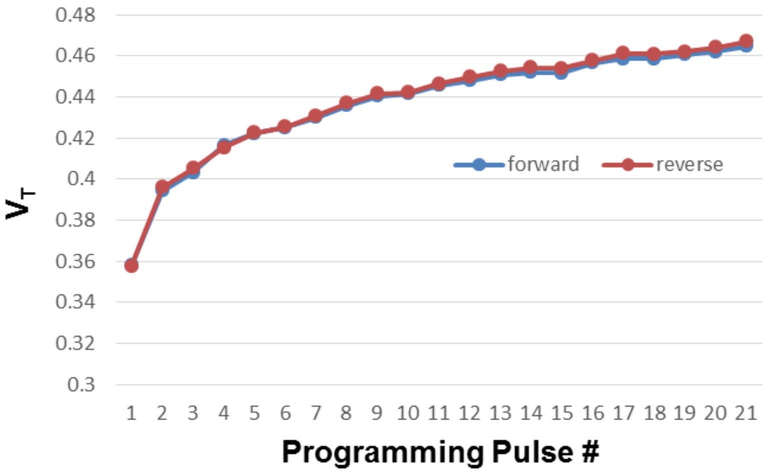


Fig. 3.15. Forward- and reverse-mode V_T measured during a programming operation for 14 nm FinFET CTTs.

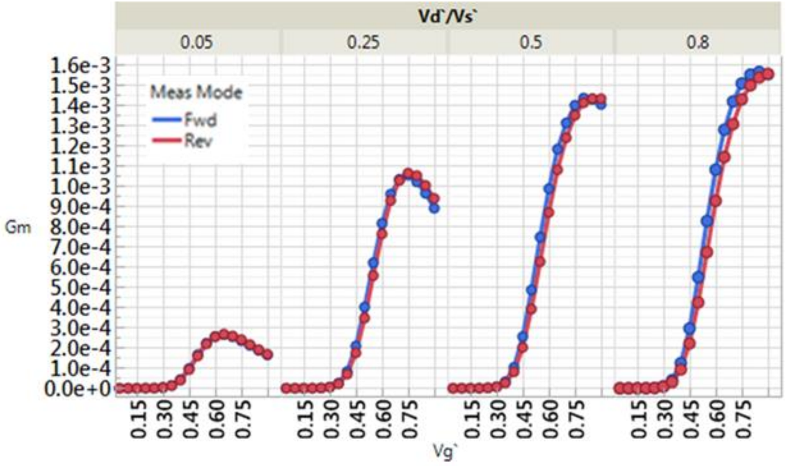


Fig. 3.16. Device transconductance vs. V_G (at various V_D/V_S values) for forward- and reverse-mode reads for 14 nm FinFET CTTs.

3.4 Data Retention

To evaluate the charge retention behavior, a set of identical 22 nm planar SOI CTTs is programmed at various fixed values of V_D to achieve a cumulative ΔV_T of ~250mV in each device and then stored at an elevated temperature of 85 °C. Retention of the trapped charge in each of the devices is measured by monitoring the device V_T's as a function of time. The reduction in V_T's (loss of trapped charge) is plotted as a percentage of the initial values as shown in Fig. 3.17. It is observed that retention of the trapped charge shows a positive correlation to the programming drain bias, V_D.

Another set of devices with different channel widths (same length) and different channel lengths (same width) is programmed using PVRS at V_D=1.5V to achieve a cumulative ΔV_T of ~265mV in each device and then stored at 85 °C. The retention of the trapped charge is measured as described above and is shown in Fig. 3.18. As can be seen, the trapped charge in wider and shorter devices has higher retention. The enhanced charge retention in wider devices is attributed to higher self-heating. The enhanced charge retention in shorter devices is attributed to a cumulative effect of higher self-heating due to higher power densities as discussed below and elevated levels of hot carriers due to higher lateral fields.

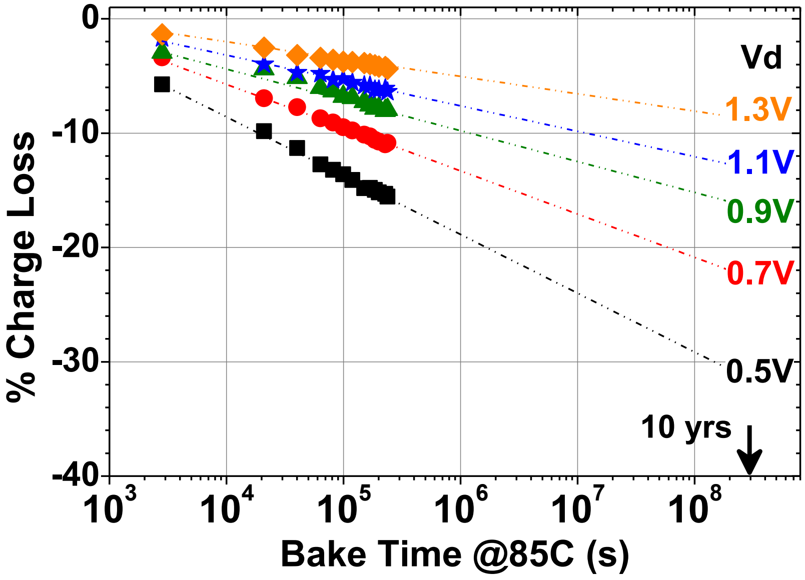


Fig. 3.17. Percentage charge loss vs. bake time @ 85 °C, for identical devices programmed at various fixed drain biases (W_{ch}=1.2 um, L=20 nm).

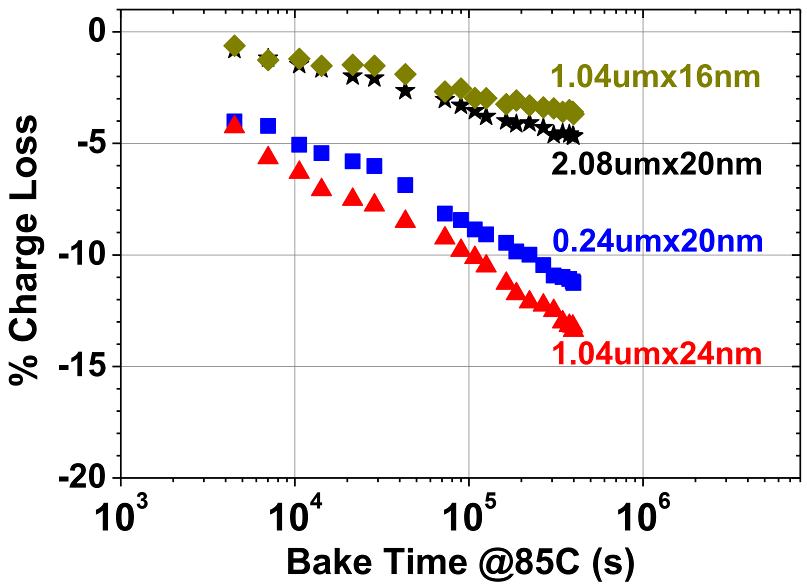


Fig. 3.18. Percentage charge loss vs. bake time @ 85 °C, for devices with various dimensions (W_{ch} x L, as labelled) programmed at V_D=1.5V.

The higher stability of charge trapped at high (device self-heating induced) temperatures, as compared to charge trapping at room temperature [9], [10], can be attributed to the fundamental nature of charge trapping and detrapping, which are thermally activated processes, wherein the capture and emission times of the trapped charge are directly correlated to their activation energies [11]. This is illustrated in Fig. 3.19. At low temperatures, stable traps with high activation energies (long capture times) require longer times to be filled (Fig. 3.19(a)). Self-heating induced high temperature enables access to these stable traps in shorter times, and they can be rapidly filled during the charge injection (Fig. 3.19(b)). Localization of self-heating leads to rapid cooling (in the ns range) after the programming conditions are removed, preventing charge detrapping as activation energies for the same can no longer be achieved, resulting in long emission times and enhanced retention (Fig. 3.19(c)). This understanding is consistent with the known properties of distributed oxide traps such as oxygen vacancies [11].

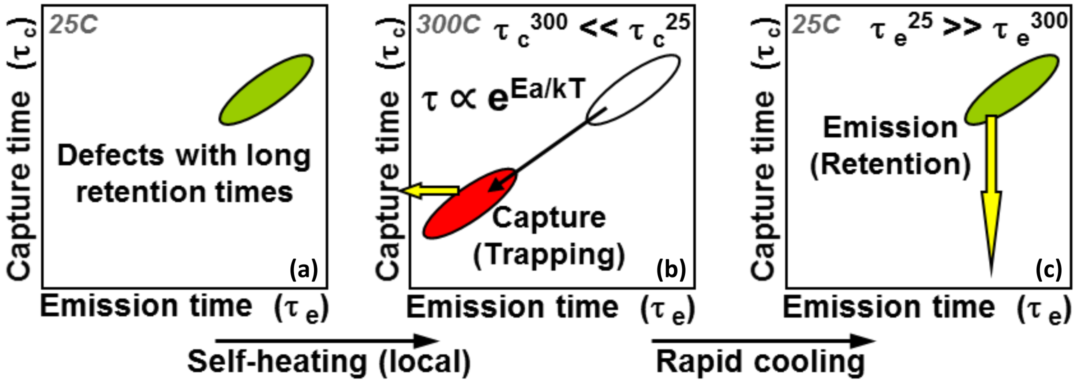


Fig. 3.19. Schematic of 'Capture-Emission Time Maps' for self-heating assisted charge trapping (adapted from [11]). (a) Defects with long emission times / good retention also have long capture times, (b) Capture times are reduced at elevated temperatures, and (c) Rapid quenching retains charge in defects with long emission times at low temperatures.

The impact of device self-heating during programming is quantified by measuring the activation energies (E_a) for charge detrapping as a function of programming V_D and self-heating temperature. The reduction in ΔV_T of devices programmed using various fixed V_D values and stored at various fixed elevated bake temperatures is monitored with time. A 'retention time' criteria of 15% ΔV_T degrade ($t_r^{15\%}$) is used and the E_a corresponding to each programming condition is extracted from an Arrhenius plot of $t_r^{15\%}$, a method commonly used in literature [10], [12]. The results (Fig. 3.20(a)) clearly show that stability of the trapped charge is significantly enhanced by programing at high V_D values (or high self-heating temperatures), consistent with all previously discussed results and speculation. This is because we are able to fill traps with higher E_a at higher programming V_D.

The existence of different types of oxygen vacancy (V_O) related electron traps in HfO_2 with various thermal activation energies for both electron trapping and detrapping has been discussed in previous literatures [9], [13], [14], [15]. The variation in capture and emission times of HfO_2 traps has also been directly correlated to the spread in their activation energies [11]. The calculated thermal activation energies for oxygen vacancy in its various charge states in crystalline *m*-HfO_2 is summarized in Fig. 3.20(b). In amorphous HfO_2, such energy levels are significantly spread in energy [11]. This is consistent with our findings in the CTT; during programming, more traps with higher activation energies for trapping are filled at higher self-heating induced temperatures (higher V_D). Such traps are also likely to be more stable, resulting in a higher effective activation energy for detrapping and enhanced stability (retention) of the CTT memory element.

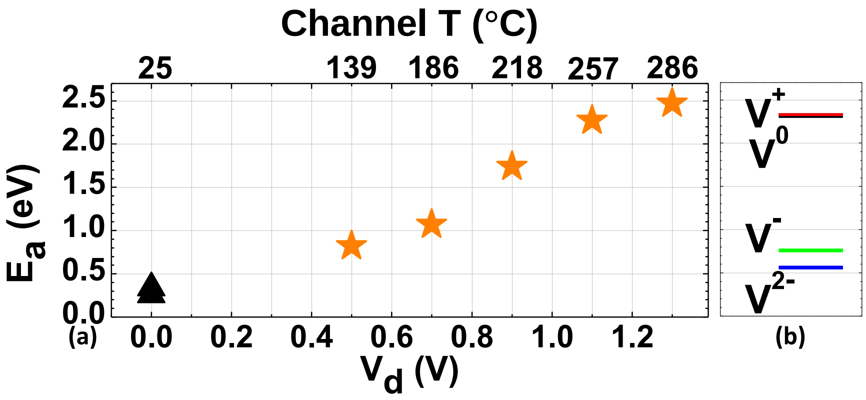


Fig. 3.20. (a) Measured activation energies (E_a) for charge detrapping after programming at various V_D values (stars). Estimated channel temperatures (in °C) due to device self-heating during programming are indicated on the top scale. Measured E_a values (triangles) for detrapping after trap filling in the absence of self-heating are shown for comparison [10]. (b) Calculated thermal activation energies for detrapping for various charge states of V_O in crystalline *m*-HfO_2 [14], revealing values ranging from 0.56eV-2.33eV.

High-temperature charge retention bake tests performed on 22 nm CTTs show a projected 10 year charge loss of <25% at 125 °C (Fig. 3.21). Results from 14 nm FinFET CTTs also show a projected charge loss of <25% after 10 years at 125 °C (Fig. 3.22).

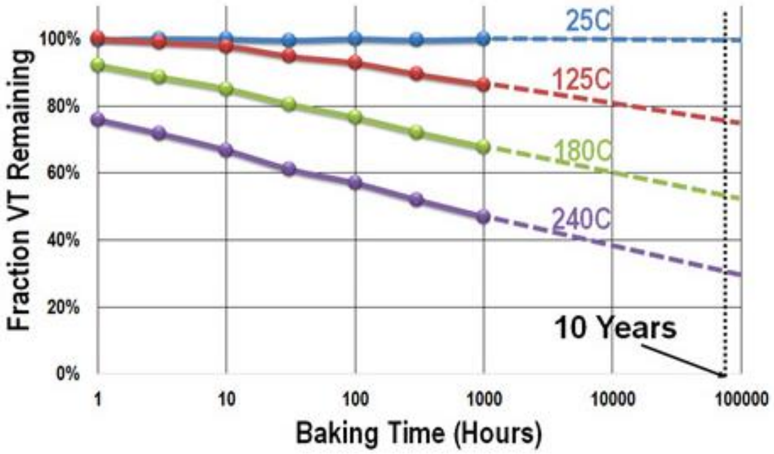


Fig. 3.21. High-temperature data retention bake tests for 22 nm SOI CTTs. Hardware results show <25% charge loss after 10 years @ 125 °C.

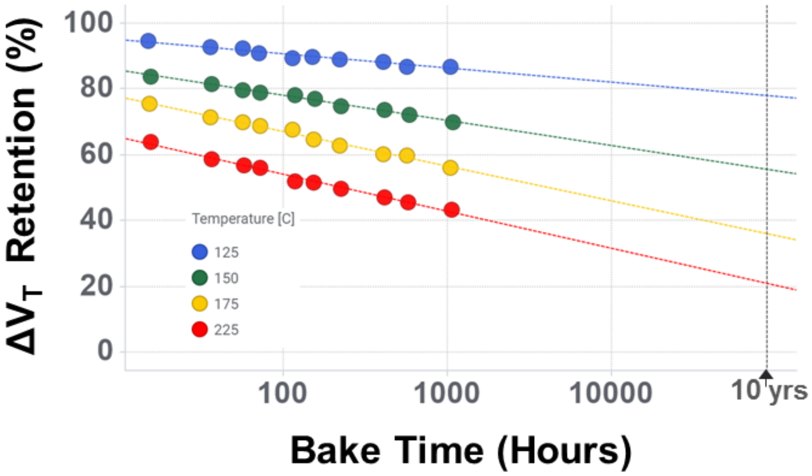


Fig. 3.22. High-temperature data retention bake tests for 14 nm FinFET CTTs programmed using V_G=2V, V_D=1.55V pulses. Hardware results show <25% charge loss after 10 years @ 125 °C. The charge detrapping activation energy (E_a), extracted using the conventional Arrhenius model, is ~1.85 eV.

The Arrhenius equation can be used to determine the acceleration factor (*AF*) as follows:

$$AF = \frac{\tau_1}{\tau_2} = e^{\left[\frac{-E_a}{k}\left(\frac{1}{T_2} - \frac{1}{T_1}\right)\right]} \tag{3.1}$$

where

T_1 = Operation temperature in Kelvin i.e. the temperature at which the memory will be operated
T_2 = Accelerated stress (bake) temperature in Kelvin
E_a = Activation energy (eV)
k = Boltzmann's constant (8.623×10^{-5} eV/K)
τ_1 = Lifetime at operation temperature
τ_2 = Lifetime at accelerated stress (bake) temperature

Once E_a has been extracted from the high-temperature bake tests (Fig. 3.22) the *AF* can been determined for a particular accelerated stress temperature (T_2) and a desired operation temperature (T_1). Once the *AF* is known, a back calculation using the Arrhenius equation (3.1) and a known reference point (T_2, τ_2) leads to fairly accurate estimates of data retention lifetimes for any operation temperature. Data retention lifetime projections for several operation temperatures, calculated using the above described method and hardware results from 14 nm FinFET CTTs, are shown in Fig. 3.23.

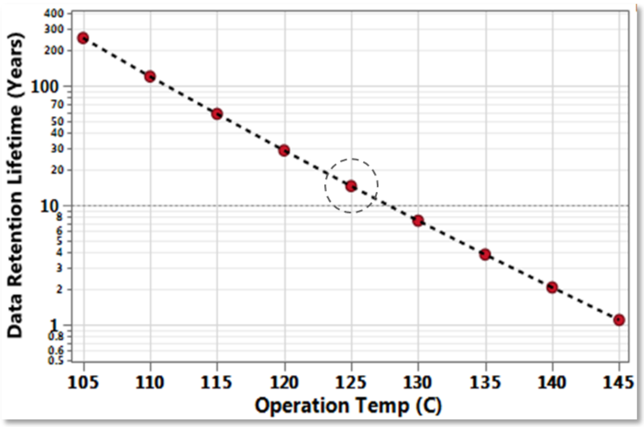


Fig. 3.23. 14 nm FinFET CTT data retention lifetime projections for several operation temperatures.

3.5 CTT OTPM Bitcell Architecture and Array Operation

In this work, a twin-cell architecture is used for the CTT OTPM (one-time programmable memory) bitcell where the data is stored on one transistor and read against an identical reference transistor. Fig. 3.24 shows a schematic of the CTT twin-cell architecture. For comparison, also shown in Fig. 3.24 are schematics of standard SRAM and DRAM bitcells. For programming a "1" in the CTT OTPM bitcell, the device corresponding to the "true bitline" (BLt) is programmed and, conversely, for programming a "0" in the CTT OTPM bitcell, the device corresponding to the "complementary bitline" (BLc) is programmed. After programming, the data is read by a common sense amplifier on each pair of bitlines. Schematics of the CTT OPTM bitcell in the 'standby', 'program', and 'read' modes along with nominal operation conditions are shown in Fig. 3.25. A universal reference, instead of the twin-cell approach, can further increase the CTT OTPM array density and is something that is under development.

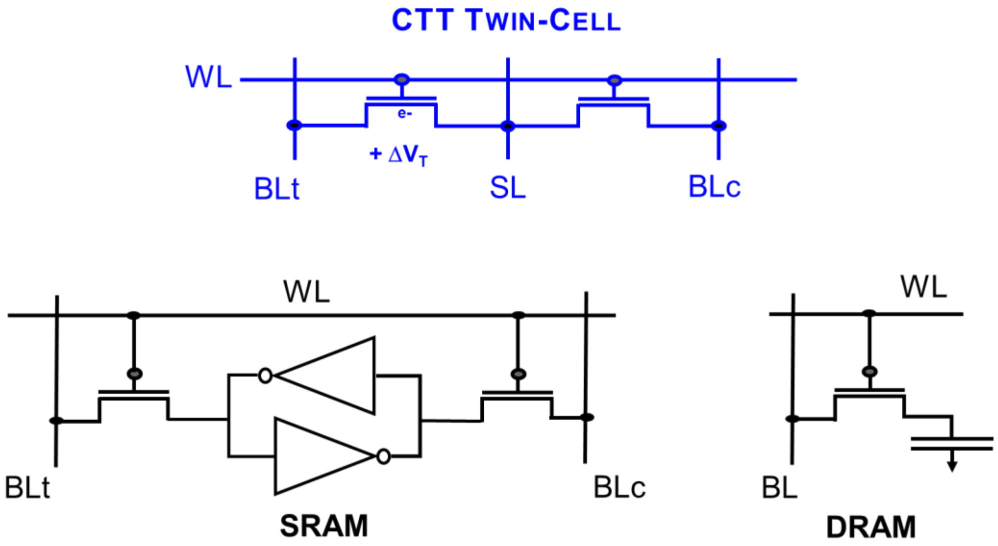
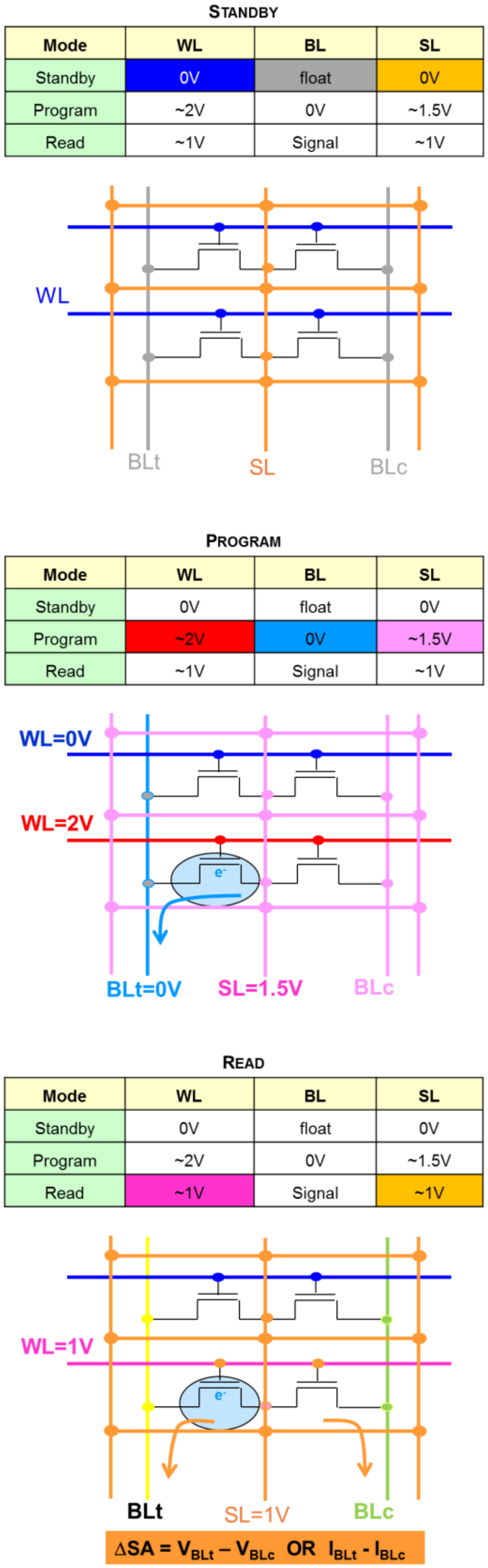


Fig. 3.24. Schematic of the CTT twin-cell. Schematics of standard SRAM and DRAM bitcells are also shown for comparison.



STANDBY

Mode	WL	BL	SL
Standby	0V	float	0V
Program	~2V	0V	~1.5V
Read	~1V	Signal	~1V

PROGRAM

Mode	WL	BL	SL
Standby	0V	float	0V
Program	~2V	0V	~1.5V
Read	~1V	Signal	~1V

READ

Mode	WL	BL	SL
Standby	0V	float	0V
Program	~2V	0V	~1.5V
Read	~1V	Signal	~1V

Fig. 3.25. Schematics of the CTT OPTM bitcell in 'standby', 'program', and 'read' modes along with nominal operation conditions.

By implementing the fundamental understanding and principles of operation of the CTT presented in this chapter, a commercially available CTT one-time programmable memory (OTPM) product designed and manufactured in a 14 nm FinFET technology, capable of operating at military grade temperatures, has already been deployed (Fig. 3.26). Circuit design aspects, including a Differential Current Sense Amplifier (DCSA) used during reads and for margining the V_T shifts during programming, are discussed in [16].

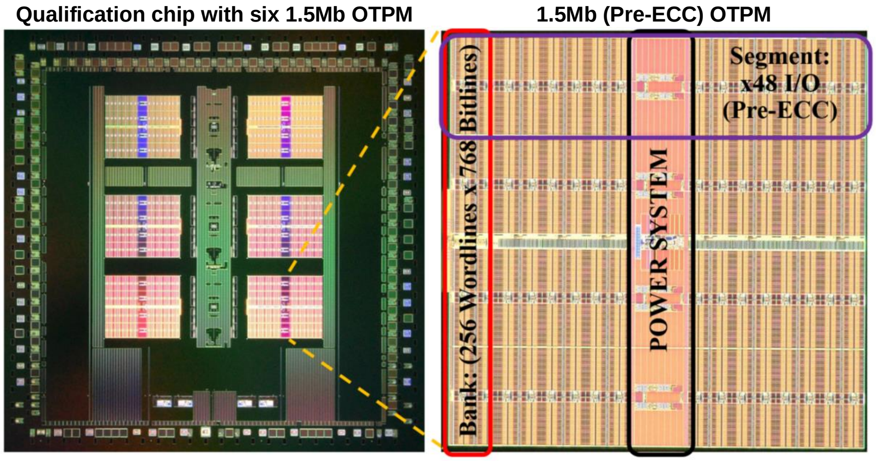

Qualification chip with six 1.5Mb OTPM
1.5Mb (Pre-ECC) OTPM
Bank: (256 Wordlines x 768 Bitlines)
POWER SYSTEM
Segment:
x48 I/O
(Pre-ECC)

Fig. 3.26. 9Mb (6×1.5Mb) CTT OTPM qualification chip photomicrographs.

Shown in Fig. 3.27 are bitmaps of a 14 nm FinFET CTT OTPM array in its native state (unprogrammed) followed by a programmed state where a checkerboard pattern has been written.

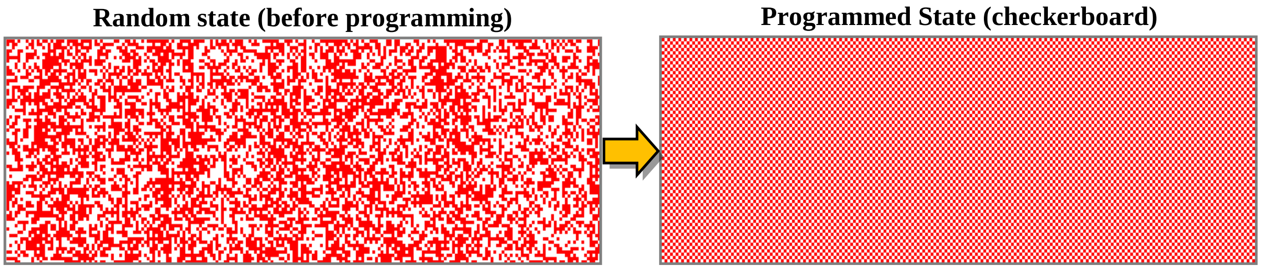


Fig. 3.27. Bitmaps of a 14 nm FinFET CTT OTPM array in its native state (unprogrammed) followed by a programmed state where a checkerboard pattern has been written.

3.6 SUMMARY

In this chapter, the fundamental understanding of self-heating enhanced charge trapping in HKMG CMOS transistors and the corresponding implications for memory applications have been presented. For HKMG CMOS transistors used as memory elements (dubbed “Charge Trap Transistors” or “CTT”) it is demonstrated that not only the magnitude but also the stability (retention) of the trapped charge significantly increases with device self-heating during the charge injection process (programming operation). The same magnitude of charge trapping can be achieved in much shorter times and/or with lower gate bias and has higher stability (retention) when the devices are programmed at higher self-heating conditions. Also presented are techniques to optimize the CTT bitcell design to enhance the programming efficiency. In particular, how device layout can be manipulated to maximize self-heating assisted charge trapping, the fundamental operation principle of CTTs, has been discussed.

The excellent data retention characteristics (> 10 years @ 125 °C), scalability, and logic voltage compatible operation make the CTT technology feasible for implementation as a fully integrated embedded non-volatile memory (eNVM) and a potential replacement for existing one-time programmable (OTP) memory technologies like eFUSE [17] and gate break down anti-fuse [18] for chip ID, on-chip encryption, field configurability, redundancy, repair, hardware security enhancement, yield improvement, and performance tailoring in HKMG CMOS technologies. Moreover, the CTT technology has a cost advantage over all other memory technologies as it requires no additional processes or masks.

While only programming related aspects of CTTs and their application as an OTP memory have been discussed in this chapter, CTTs can also be employed as a multi-time programmable (MTP) memory, which would of course require erasing the programmed devices efficiently. The technological breakthroughs required for implementation of CTTs as an MTP memory in 14 nm FinFET technologies and beyond, with an endurance of > 10^4 program/erase cycles, data retention of > 10 years at 125 °C, and operation capability at military grade temperatures are discussed in subsequent chapters.

REFERENCES

[1] S. Narasimha, P. Chang, C. Ortolland, D. Fried, E. Engbrecht, K. Nummy, P. Parries, T. Ando, M. Aquilino, N. Arnold, R. Bolam, J. Cai, M. Chudzik, B. Cipriany, G. Costrini, M. Dai, J. Dechene, C. DeWan, B. Engel, M. Gribelyuk, D. Guo, G. Han, N. Habib, J. Holt, D. Ioannou, B. Jagannathan, D. Jaeger, J. Johnson, W. Kong, J. Koshy, R. Krishnan, A. Kumar, M. Kumar, J. Lee, X. Li, C-H. Lin, B. Linder, S. Lucarini, N. Lustig, P. McLaughlin, K. Onishi, V. Ontalus, R. Robison, C. Sheraw, M. Stoker, A. Thomas, G. Wang, R. Wise, L. Zhuang, G. Freeman, J. Gill, E. Maciejewski, R. Malik, J. Norum and P. Agnello, “22nm High-Performance SOI Technology Featuring Dual-Embedded Stressors, Epi-Plate High-K Deep-Trench Embedded DRAM and Self-Aligned Via 15LM BEOL,” *IEEE IEDM*, 2012, pp. 3.3.1–3.3.4.

[2] A. Kerber, S. Krishnan and E. Cartier, “Voltage Ramp Stress for Bias Temperature Instability Testing of Metal-Gate/High-*k* Stacks,” *IEEE Electron Device Letters*, 30 (12), 2009.

[3] P. Su, K. Goto, T. Sugii and C. Hu, “Excess Hot-Carrier Currents in SOI MOSFETs and Its Implications,” *IEEE IRPS*, 2002, pp. 93-97.

[4] D. Dallmann and K. Shenai, “Scaling Constraints Imposed by Self-Heating in Submicron SOI MOSFET’s,” *IEEE Transactions on Electron Devices*, 42 (3), pp. 489-496, 1995.

[5] N. Raghavan, K. Pey, and K. Shubhakar, “High-κ dielectric breakdown in nanoscale logic devices – Scientific insight and technology impact,” *Microelectronics Reliability*, 54 (5), 2014, pp. 847-860.

[6] J. Singh, A. Bousquet, J. Ciavatti, K. Sundaram, J. S. Wong, K. W. Chew, A. Bandyopadhyay, S. Li, A. Bellaouar, S. M. Pandey, B. Zhu, A. Martin, C. Kyono, J.-S. Goo, H. S. Yang, A. Mehta, X. Zhang, O. Hu, S. Mahajan, E. Geiss, S. Yamaguchi S. Mittal, R. Asra, P. Balasubramaniam, J. Watts, D. Harame, R. M. Todi, S. B. Samavedam and D.K. Sohn, "14nm FinFET Technology for Analog and RF Applications," *IEEE Symp. VLSI Technol.*, 2017, pp. T140-T141.

[7] F. Crupi, R. Degraeve, A. Kerber, D.H. Kwak and G. Groeseneken, "Correlation between Stress-Induced Leakage Current (SILC) and the HfO_2 Bulk Trap Density in a SiO_2 / HfO_2 Stack," *IEEE IRPS*, 2004.

[8] Y.-P. Gong, A.-D. Li, X. Qian, C. Zhao and D. Wu, "Interfacial structure and electrical properties of ultrathin HfO_2 dielectric films on Si substrates by surface sol–gel method," *J. Phys. D: Appl. Phys.*, 42 (1), 2009.

[9] E. Cartier, B. P. Linder, V. Narayanan and V. K. Paruchuri, "Fundamental understanding and optimization of PBTI in nFETs with SiO_2/HfO_2 gate stack," *IEEE IEDM*, 2006, pp. 1-4.

[10] G. Bersuker, J. H. Sim, C. S. Park, C. D. Young, S. V. Nadkarni, R. Choi and B. H. Lee, "Mechanism of Electron Trapping and Characteristics of Traps in HfO_2 Gate Stacks," *IEEE Trans. Device Mater. Rel.*, vol. 7, no.1, pp.138-145, 2007.

[11] T. Grasser, P.-J. Wagner, H. Reisinger, Th. Aichinger, G. Pobegen, M. Nelhiebel and B. Kaczer, "Analytic modeling of the bias temperature instability using capture/emission time maps," *IEEE IEDM*, 2011, pp. 27.4.1-27.4.4.

[12] G. Molas, M. Bocquet, E. Vianello, L. Perniola, H. Grampeix, J.P. Colonna, L. Masarotto, F. Martin, P. Brianceau, M. Gély, C. Bongiorno, S. Lombardo, G. Pananakakis, G. Ghibaudo and B. De Salvo, "Reliability of charge trapping memories with high-*k* control dielectrics," *Microelectronic Engineering*, vol. 86, pp. 1796-1803, 2009.

[13] E. P. Gusev and C. P. D'Emic, "Charge detrapping in HfO_2 high-*κ* gate dielectric stacks," *Applied Physics Letters*, vol. 83, no. 25, pp.5223-5225, 2003.

[14] J. L. Gavartin, D. M. Ramo, A. L. Shluger, G. Bersuker and B. H. Lee, "Negative oxygen vacancies in HfO_2 as charge traps in high-*k* stacks," *Applied Physics Letters*, vol. 89, 2006.

[15] K. Xiong, J. Robertson, M. C. Gibson and S. J. Clark, "Defect energy levels in HfO_2 high-dielectric-constant gate oxide," *Applied Physics Letters*, vol. 87, 2005.

[16] E. Hunt-Schroeder, D. Anand, J. Fifield, M. Roberge, D. Pontius, M. Jacunski, K. Batson, M. Deming, F. Khan, D. Moy, A. Cestero, R. Katz, Z. Chbili, E. Banghart, L. Jiang, B. Jayaraman, R. R. Tummuru, R. Raghavan, A. Mishra, N. Robson and T. Kirihata, "14nm FinFET 1.5Mb Embedded High-K Charge Trap Transistor One Time Programmable Memory Using Differential Current Sensing," *IEEE Solid-State Circuits Letters*, vol. 1, no. 1, 2019.

[17] C. Kothandaraman, S. K. Iyer and S. S. Iyer, "Electrically Programmable Fuse (eFUSE) Using Electromigration in Silicides," *IEEE Electron Device Letters*, 23 (9), 2002, pp. 523–525.

[18] Y. Liu, M.H. Chi, A. Mittal, G. Aluri, S. Uppal, P. Paliwoda, E. Banghart, K. Korablev, B. Liu, M. Nam, M. Eller and S. Samavedam, "Anti-Fuse Memory Array Embedded in 14nm FinFET CMOS with Novel Selector-Less Bit-Cell Featuring Self-Rectifying Characteristics," *IEEE Symp. VLSI Technol. Dig. Tech. Papers*, 2014.

4. CHARGE TRAP TRANSISTORS (CTT) AS A MULTI-TIME PROGRAMMABLE EMBEDDED NON-VOLATILE MEMORY

In the previous chapter, it was demonstrated how intrinsic self-heating enhanced charge trapping can be exploited in HKMG devices to achieve large and stable threshold voltage (V_T) shifts that are suitable for non-volatile memory applications. In this chapter, it is demonstrated that indeed multi-time programmability is possible for application of CTTs as a multi-time programmable memory (MTPM) technology. The underlying principles of operation, key factors for operation optimization, challenges, and corresponding solutions are presented.

4.1 PRINCIPLES OF OPERATION

A schematic of the basic operation of a CTT memory device is depicted in Fig. 4.1; the device V_T is modulated by the charge trapped in the high-*k* dielectric of the HKMG device. The reader is reminded that, while schematics of a planar device are shown here for demonstration, the same fundamental principles equally apply to FinFET based CTTs, as demonstrated and discussed in detail in Chapter 3.

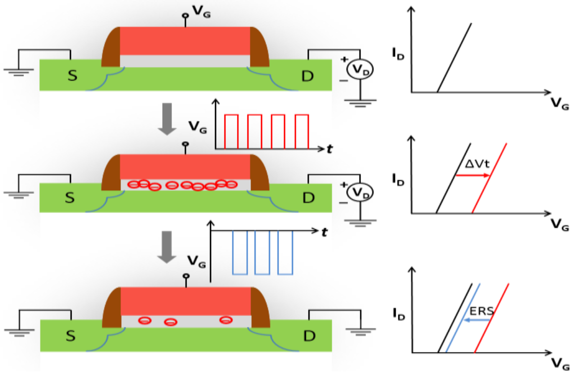


Fig. 4.1. A schematic depicting the basic operation of a CTT memory device (equally applicable to planar FET as well as FinFET based CTTs).

In order to understand the dynamic behavior of charge trapping in CTTs, device V_T shifts (ΔV_T) are first measured as a function of the programming time (t_P), as shown in Fig. 4.2(a). 1.2µm×20nm devices (22nm SOI technology [1]) are programmed using gate voltage (V_G) pulses with a magnitude of 2V while the drain-to-source voltage (V_D) is fixed at 1.3V. ΔV_T vs. stress time with 2V V_G pulses and V_D=0V (I_{DS}=0) is also shown for comparison between self-heating enhanced charge trapping at high V_D [2] and conventional Positive Bias Temperature Instability (PBTI) [3] where there is no channel current flowing and hence no self-heating is present. It is observed that ΔV_T is dramatically enhanced when the transistor is pulsed at high V_D [2] and it shows a logarithmic dependence on t_P; Programming efficiency is highest at the beginning of the program operation and reduces with increasing programming time as more and more of the available electron traps are filled. The measured peak power during the program operation is ~4mW (Fig. 4.2(b)), which is considerably less than the typical power required to program an eFUSE in the same technology (~20mW), allowing us to use smaller driver transistors to achieve programming, compared to the eFUSE case. Peak eFUSE power does not scale appreciably and can even increase significantly as more refractory metals are used as fuse elements. Fig. 4.2(c) shows the calculated energy (E_P=$\int I_D \times V_D \times t_P$) required vs. the measured ΔV_T achieved. As can be seen, as ΔV_T increases the energy required to create any additional V_T shift increases rapidly, which reinforces the message being conveyed by Fig. 4.2(a) i.e. programming efficiency reduces as t_P increases.

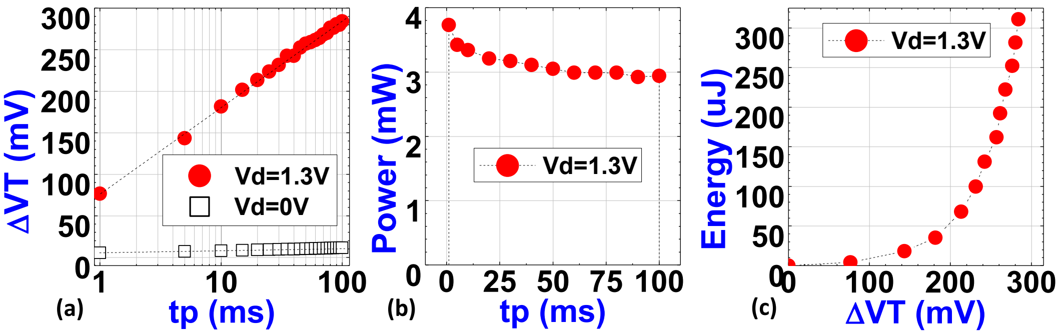


Fig. 4.2. (a) ΔV_T vs. t_P (V_G =2V, V_D =1.3V). ΔV_T for PBTI @ 25°C (V_G =2V, V_D =0V) is shown for comparison. (b) Power consumption vs. time during programming. (c) Total energy (E_p=$\int I_D \times V_D \times t_P$) required vs. target ΔV_T.

To understand the Program/Erase (P/E) characteristics and the fundamental physical mechanisms behind the operation of CTT memory devices, P/E cycling of the devices is performed using the Pulsed gate Voltage Ramp Sweep (PVRS) technique (details on PVRS discussed in [2] and [4]), with 10ms V_G pulses of increasing magnitudes in 10mV increments, as demonstrated in Fig. 4.3, for various fixed programming V_D values. The very first program operation, referred to as 'initialization', is unique. This is followed by an erase ('ERS') operation using negative PVRS and then a re-program ('PRG') operation. The source and drain are typically grounded during the erase operations. The observed behavior reveals the presence of three distinct V_D-dependencies which can be exploited in a CTT for an MTPM application; *(i)* As seen during 'initialization', ΔV_T has a strong V_D-dependence; At higher V_D, equivalent ΔV_T values are achievable at much lower V_G. This effect is due to a combination of enhanced trapping and trap creation in the HfO_2 at higher V_D (stronger device self-heating) as discussed in [2], [5], [6], [7] and in more detail below. *(ii)* For devices programmed at higher V_D, longer times and/or larger negative V_G values are needed to de-trap the charge. Charge trapping at high temperature (stronger self-heating at high V_D) is more stable and it is more difficult to erase the devices. This is consistent with what was reported in [2], where enhanced charge retention was demonstrated for devices programmed at higher V_D. The slight ΔV_T difference between the end of the 'initialization' cycle and beginning of the 'ERS' cycle is believed to be caused by fast de-trapping of the small fraction of unstable trapped charge in each case, followed by no further de-trapping until a certain negative bias is applied during the 'ERS' cycle. The magnitude of this small ΔV_T difference is inversely proportional to the programming V_D, which is again consistent with the relation between programming V_D and overall trapped charged stability, as discussed in Chapter 3. *(iii)* The charge trapping behavior changes after the 'initialization' operation: this is due to the creation of new traps [5], [6], allowing for subsequent programming ('PRG') to the same ΔV_T at lower V_G. This phenomenon has also been reported in [8], where an increased rate of charge trapping for pre-stressed devices is attributed to new trap creation during the charge injection process. In order to verify the above and compare self-heating enhanced charge trapping to conventional BTI, PVRS sweeps were also done with V_D=0V (Fig. 4.3 inset). It is clearly seen that, without the effect of self-heating, *(i)* for the same V_G values, the ΔV_T achieved is relatively very small, *(ii)* the ΔV_T is fully recoverable (i.e. traps discharge easily), and most importantly *(iii)* the charge trapping behavior does not change subsequent to the first cycle and is repeatable for many cycles, indicating that creation of additional traps is minimal. These findings regarding the impact of device self-heating on the magnitude and stability of ΔV_T are consistent with the findings and conclusions presented in Chapter 3.

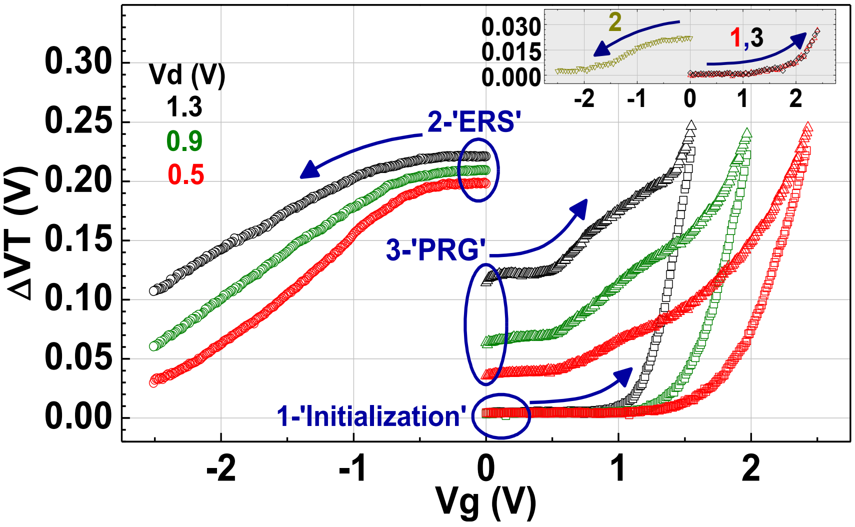


Fig. 4.3. Measured ΔV_T during 1-‘Initialization’, 2-’ERS’, and 3-’PRG’ cycles for various V_D values using PVRS. Inset shows ΔV_T for V_D =0V PVRS stress (BTI).

There is an obvious trade-off between trapped charge retention, the ΔV_T window, and the erase time/voltage needed; Higher programming V_D results in more stable V_T shifts (better retention), as demonstrated and discussed in detail in Chapter 3, but it will take longer time and/or higher voltage to erase the cells. In other words, for a given erase time/voltage constraint, the ΔV_T window will be smaller if higher programming V_D is used. Therefore, it is important to optimize the operating conditions of the memory cells. Typically, erase times longer than programming times are needed to avoid under-erasing and programming times shorter than those in the ‘initialization’ operation are needed to avoid over-programming, in order to achieve a sufficiently large memory window. It is also advantageous to perform ‘initialization’ at a higher V_D than that subsequently used during the ‘PRG’ operations to avoid over-programming in subsequent P/E cycles. At the same time, V_D for programming must be selected high enough for the trapped charge to have acceptable retention for the memory application. While this discussion provides a general guideline, detailed optimization will depend on device geometry, layout, and gate stack properties, all of which affect the charge trapping behavior [2], [7].

4.2 PROGRAM AND ERASE OPTIMIZATION AND CYCLING

To demonstrate the importance of optimizing the program and erase conditions for the CTT MTPM, as discussed in the previous section, devices are cycled 20× using unoptimized P/E conditions (i.e. P/E conditions are not optimized to avoid over-programming and under-erasing) as well as optimized P/E conditions ($V_{D\text{-}PRG}$ slightly lower than $V_{D\text{-}INIT}$ is used and the number of 'PRG' pulses is limited to avoid over-programming. Longer erase times are used during the 'ERS' operation to achieve maximum ΔV_T recovery) for comparison. Post-program and post-erase ΔV_T values for the devices in each case are shown in Fig. 4.4(a). It is clear that, by optimizing P/E conditions, over-programming and under-erasing with P/E cycling (which causes the memory window to dynamically drift higher, resulting in a shrinking read-margin for the "erased" state with respect to a fixed reference read voltage, as seen with unoptimized P/E conditions) can be avoided, resulting in significant improvement in the endurance of the memory cells. In functional memory arrays, program and erase 'verify' schemes are used to further optimize the P/E operations. Fig. 4.4(b) shows the post-program and post-erase ΔV_T values for devices that were cycled 800× using optimized P/E conditions. It can be seen that, even after 800 cycles, a stable memory window (~120mV in this case) exists.

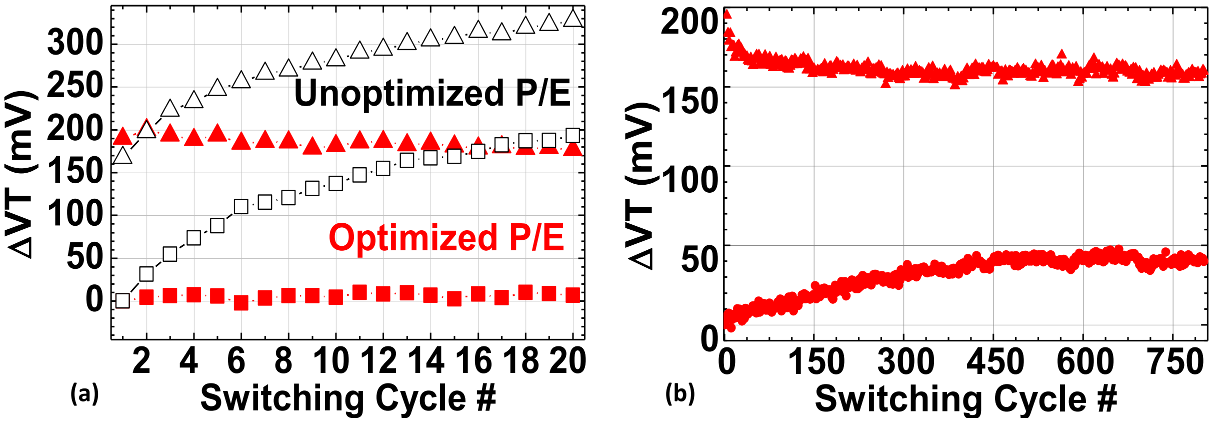


Fig. 4.4 (a) Memory window vs. switching cycle number comparison between un-optimized P/E ($V_{G\text{-}INIT}$=2V, $V_{D\text{-}INIT}$=1.3V, $V_{G\text{-}PRG}$=2V, $V_{D\text{-}PRG}$=1.3V, $V_{G\text{-}ERS}$=-2V, open black symbols) and optimized ($V_{G\text{-}INIT}$=2V, $V_{D\text{-}INIT}$=1.3V, $V_{G\text{-}PRG}$=2V, $V_{D\text{-}PRG}$=1.2V, $V_{G\text{-}ERS}$=-2V, solid red symbols) P/E conditions. (b) 800× P/E cycles using optimized P/E conditions.

While the ‘initialization’ technique is a cornerstone that enables the implementation of CTTs as an MTPM, there is still a need for improvement to the erase efficiency. As seen in Fig. 4.4(b), while the memory window narrowing can be significantly mitigated using the ‘initialization’ technique, there is nonetheless a narrowing of the window. Even if the post-program V_T is very accurately controlled using program verify techniques, a drift in post-erase V_T is still observed. This becomes particularly troublesome as devices are further scaled down and the ΔV_T windows correspondingly shrink: when the ΔV_T window is small to begin with, even a small narrowing significantly reduces the number of P/E cycles before the memory window is pinched-off, hence significantly reducing the P/E cycling endurance the memory can offer. Additionally, it is observed that the erase efficiency in FinFET CTTs is lower as compared to planar CTTs: this is likely due to the non-uniform electric-field distribution in the fin structure. This effect is demonstrated on 14nm bulk FinFET CTTs (Fig. 4.5). Shown in Fig. 4.5(a) is P/E cycling with unoptimized P/E conditions. As expected, a drift in both post-program and post-erase sense currents (I_{SENSE}) is seen and the memory window is seen to collapse in ~10 cycles. Shown in Fig. 4.5(b) is P/E cycling with optimized programming conditions (determined using the 'initialization' technique) and program verify to avoid over-programming. It is clear that even when the post-program I_{SENSE} is kept fairly constant, the drift in the post-erase I_{SENSE} due to the under-erase after each cycle causes the memory window to pinch-off in ~15 P/E cycles or so. It is clear that a different/better erase technique is needed for CTTs in 14nm FinFET technologies (and beyond). Indeed, poor erase efficiency - and the consequent low P/E cycling endurance - has restricted implementation of the CTT technology as an MTPM in the 14nm FinFET node thus far.

In order to address the erase problem, introduced is a technique called "Self-heating Temperature Assisted eRase" (STAR) [9], [10], [11], [12] that dramatically improves the erase efficiency, and in turn, the cycling endurance of the CTT MTPM. For the first time, hardware results demonstrate an endurance of > 10^4 P/E cycles, a 1000× improvement, which is adequate for most embedded MTPM applications such as hardware security, encryption, firmware, configuration, and repair.

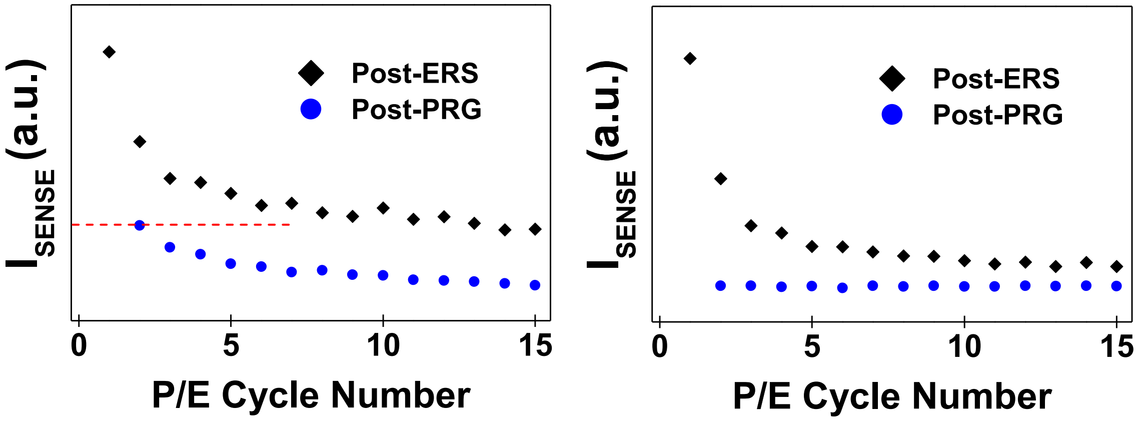


Fig. 4.5. 14nm FinFET CTT P/E cycling using (a) Unoptimized conditions and (b) optimized programming conditions with 'initialization' and program verify.

4.3 "Self-heating Temperature Assisted eRase" (STAR)

Conventional erase operations (Fig. 4.6(a)), typically performed using a negative gate bias (V_G) of magnitude > |2.5V|, while the source, drain and substrate are grounded, to electrostatically emit trapped charge, result in an inefficient erase (Fig. 4.7(a)). Higher voltages cannot be used due to gate oxide breakdown concerns. The incomplete erase after each cycle causes the memory window to dynamically drift and become narrower, resulting in a shrinking read margin (Fig. 4.8(a)). This severely limits the endurance (< 15 P/E cycles) and makes it challenging for implementation of CTTs as an MTPM technology, as circuits to dynamically change the reference current are difficult to implement. This problem is effectively addressed by using the STAR technique. Charge de-trapping in high-*k* dielectrics is strongly accelerated by temperature, usually defined by the Arrhenius model. The STAR technique (Fig. 4.6(b)), utilizes the source-substrate-drain structure of the device as a parasitic NPN bipolar junction transistor (BJT) to pass a short current pulse through the body of the device during the erase operation. The device is biased such that, the parasitic BJT is in the active mode while there is a negative gate-to-substrate bias (V_{GX}) at the same time, without the need for any negative voltages. The local device self-heating caused by the BJT current, in combination with the negative V_{GX}, significantly enhances the charge de-trapping process: up to 100% erase efficiency (Fig. 4.7(b)) is achieved using lower voltages and shorter time as compared to the conventional erase method (100% erase within 1ms using STAR vs. < 50% even after 1s of conventional erase), in turn yielding a flat memory window with no narrowing for 10^4 P/E cycles (Fig. 4.8(b)). 3D finite element thermal simulations of the respective bitcell temperatures during the erase operations performed using the two methods are also shown in Fig. 4.6(a) and Fig. 4.6(b). Simulation results estimate that steady state T is achieved within ~40-50 ns (Fig. 4.9). The measured I-V characteristics of the parasitic BJT are shown in Fig. 4.10.

The five-transistor STAR enabled MTPM bitcell design is shown in Fig. 4.11. In the memory array, each bitcell receives nine wires used to control or supply voltages generated on-chip from a 2.5V power supply during the modes of operation, shown in Fig. 4.11. The array is partitioned such that each wordline has a dedicated source line domain, which isolates the erase disturb (charge loss) to the bitcells on a common wordline. The bitcells that are exposed to the ~2V V_X and V_D (i.e. bitlines / columns on the same wordline) are sequentially erased due to this charge loss condition. However, cells on adjacent wordlines maintain a grounded V_X and V_D thereby avoiding erase disturb. The erase disturb isolation is done using four pFETs (Fig. 4.11) passing the voltages only to the row of memory cells that need to be erased. The bitcell pFETs are area efficient thin oxide devices, requiring stacking to keep transistors in safe operating regions through all modes of operation.

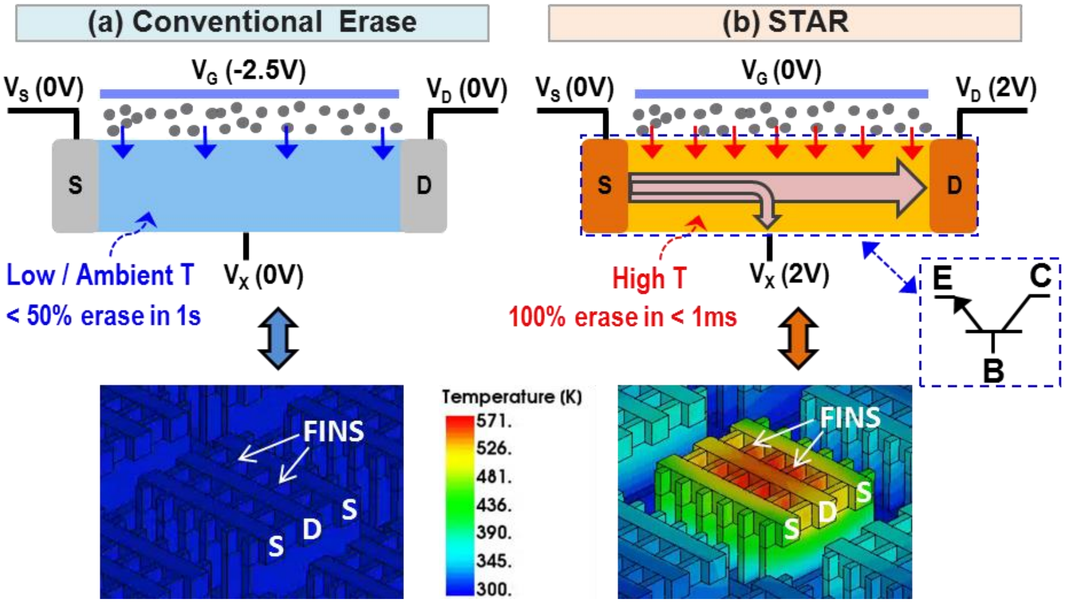


Fig. 4.6. Schematic showing (a) conventional erase and (b) “Self-heating Temperature Assisted eRase” (STAR). Corresponding thermal profiles of the bitcells during the erase operations are also shown for comparison.

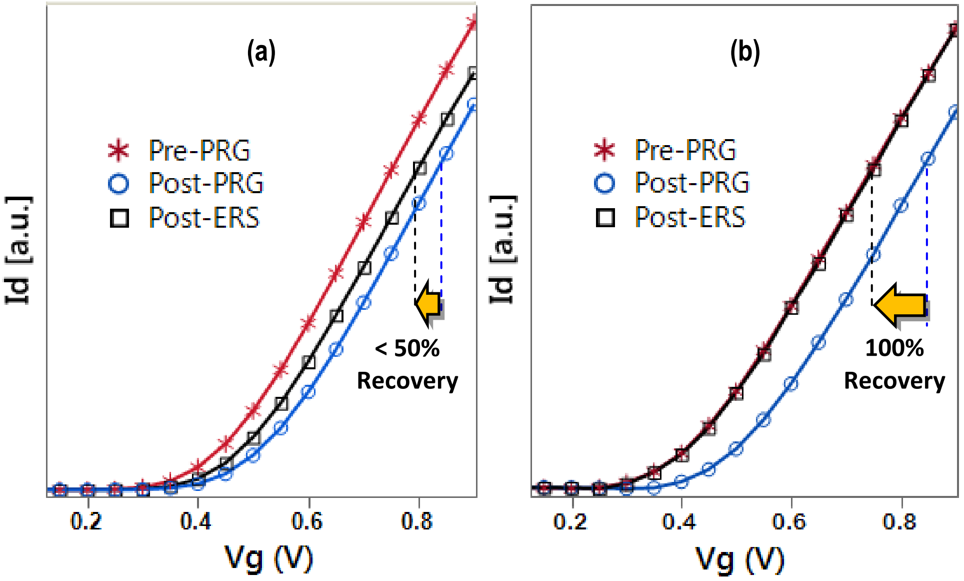


Fig. 4.7. Measured Pre-Program, Post-Program, and Post-Erase I_D-V_G data with (a) conventional erase and (b) STAR.

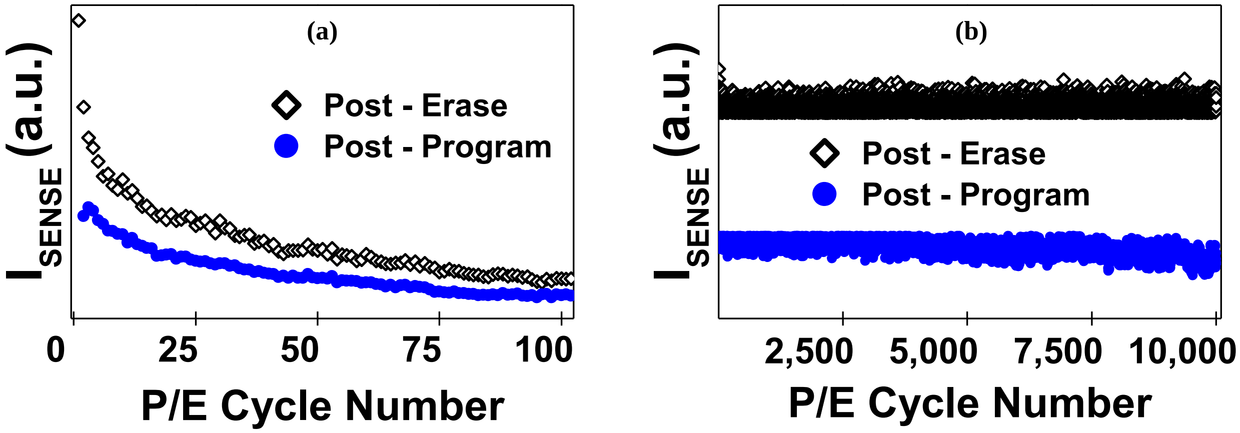


Fig. 4.8. P/E cycling of 14 nm FinFET CTTs using (a) conventional erase and (b) STAR.

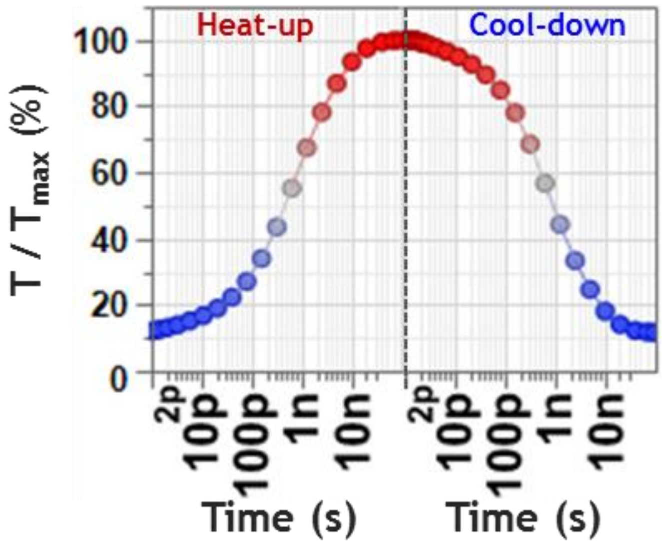


Fig. 4.9. Transient bitcell temperature vs. time for self-heating (during erase) and subsequent cool-down.

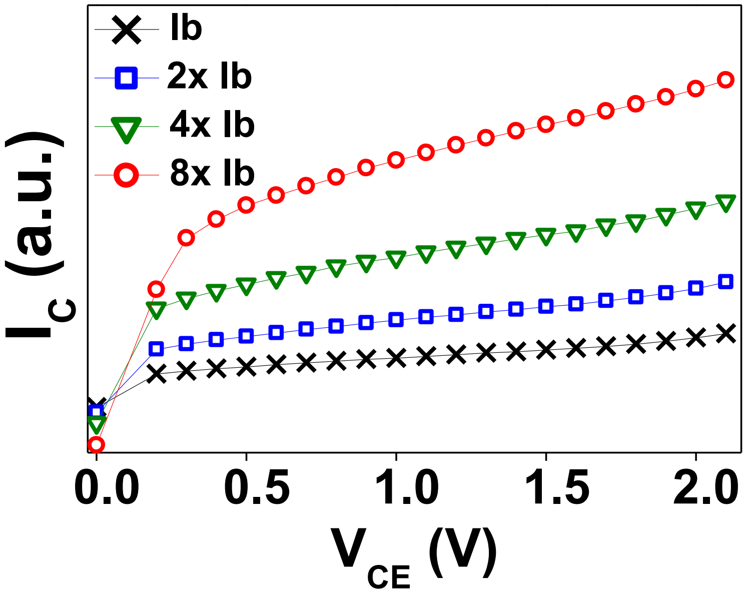
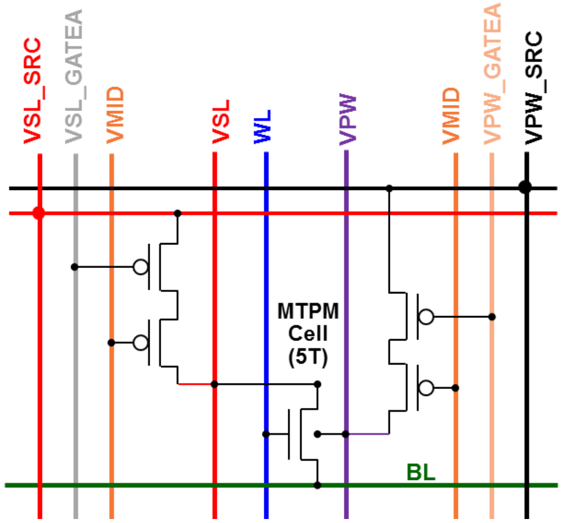


Fig. 4.10. Measured I-V characteristics of the parasitic BJT (described in Fig. 4.6(b)).



Signal Voltages for each mode	Standby	Write Mask	Write Select	Erase Mask	Erase Select	Read
Voltages						
VSL_SRC	0	1.6	1.6	2	2	0
VSL	0	0	1.6	0	2	0
VPW_SRC	0	0	0	2	2	0
VPW	0	0	0	0	2	0
VMID	0.67	0.67	0.67	0.84	0.84	0.67
Gate controls						
VSL_GATEA	VDD	1.6	0.67	2	0.84	VDD
WL	0	0	1.9	0	0	~0.45
VPW_GATEA	VDD	VDD	VDD	2	0.84	VDD

Fig. 4.11. STAR enabled CTT bitcell design and typical operation conditions in 14 nm FinFET technology.

4.4 Data Retention

High-temperature charge retention bakes, performed on 14 nm FinFET CTTs cycled using V_G=1.95V, V_D=1.55V for programming and erased using STAR, show a projected 10 year charge loss of < 25% at 125 °C (Fig. 4.12). The charge de-trapping activation energy (E_a), extracted using the Arrhenius model, is ~1.85 eV. This is comparable to the charge de-trapping E_a for one-time programmable 14 nm FinFET CTTs reported in Chapter 3. Differential sense current (Δ I_{SENSE}) distributions for a 9kb CTT array baked at 125 °C, for up to seven days (168 hours), are shown in Fig. 4.13.

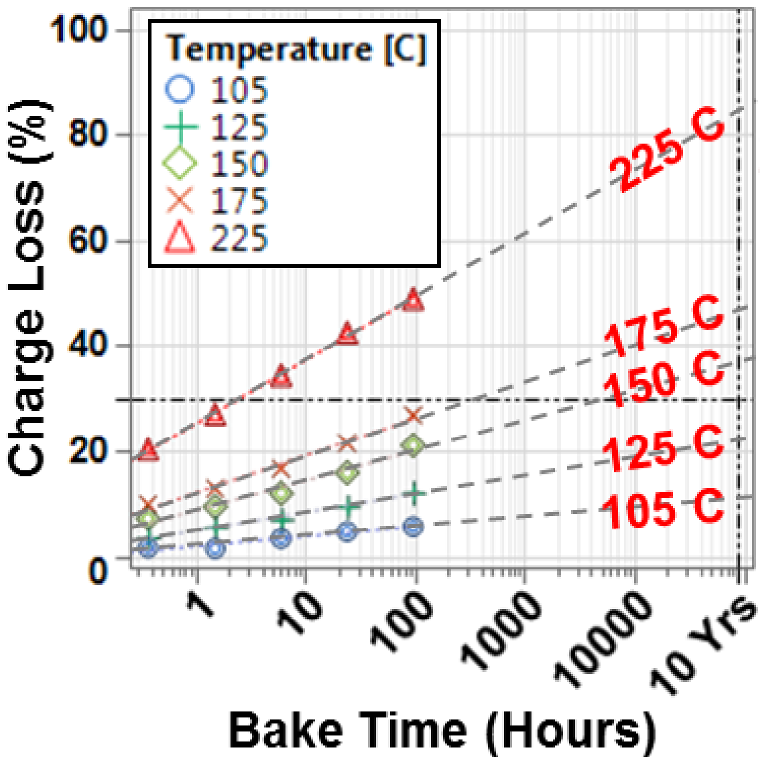


Fig. 4.12. High-temperature data retention bake tests showing a charge loss of <25% in 10 years at 125 °C.

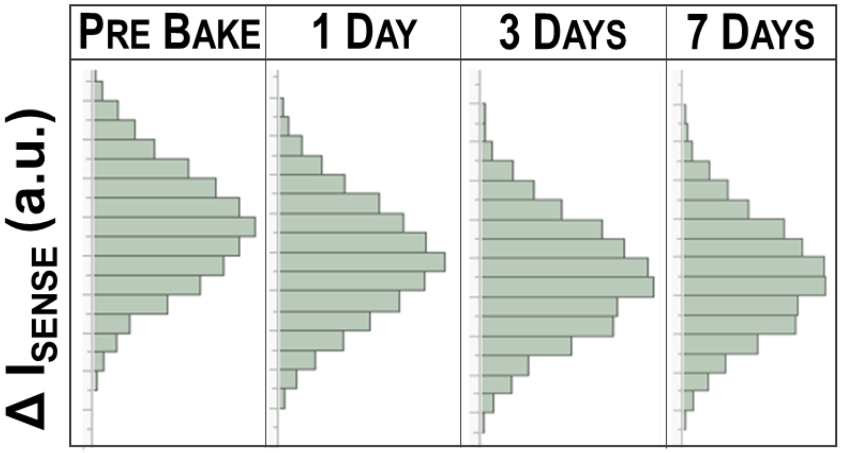


Fig. 4.13. Δ I_{SENSE} distributions in a 9kb CTT array baked at 125 °C, for up to 7 days.

A functional macro with a CTT MTPM array with STAR implementation, designed and manufactured in a 14 nm FinFET technology, is shown in Fig. 4.14. Also shown therewith are the measured sense currents for programmed and erased states during P/E cycling using STAR. A very flat memory window is achieved with low variability in post-program and post-erase sense currents. Unlike the CTT OTPM array where a twin-cell architecture is used, the CTT MTPM array is composed of single transistor bitcells with a universal reference.

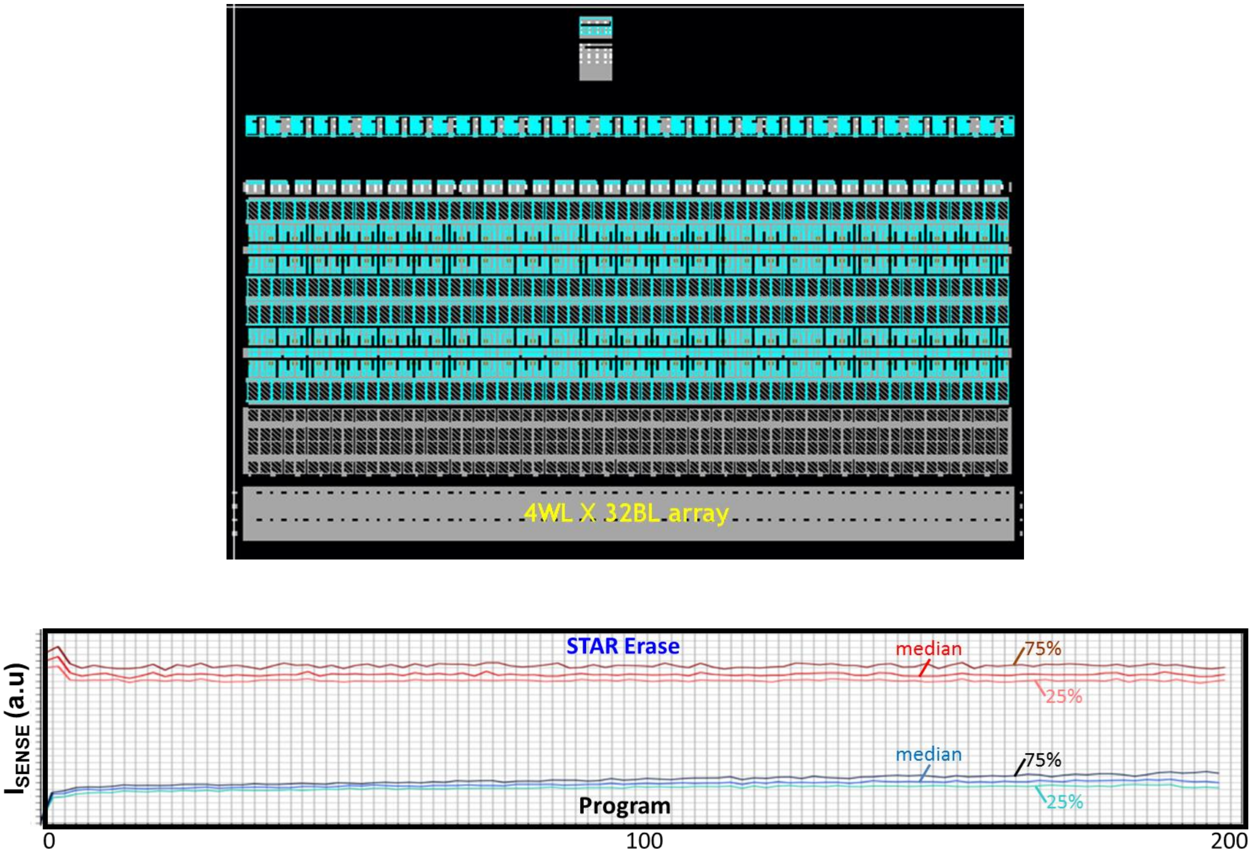


Fig. 4.14. A 14 nm FinFET CTT array and P/E cycling using STAR.

Bitmaps of CTT MTPM arrays integrated in 32 nm SOI planar, 22 nm SOI planar, 14 nm SOI FinFET, and 14 nm bulk FinFET production technologies are demonstrated in Fig. 4.15.

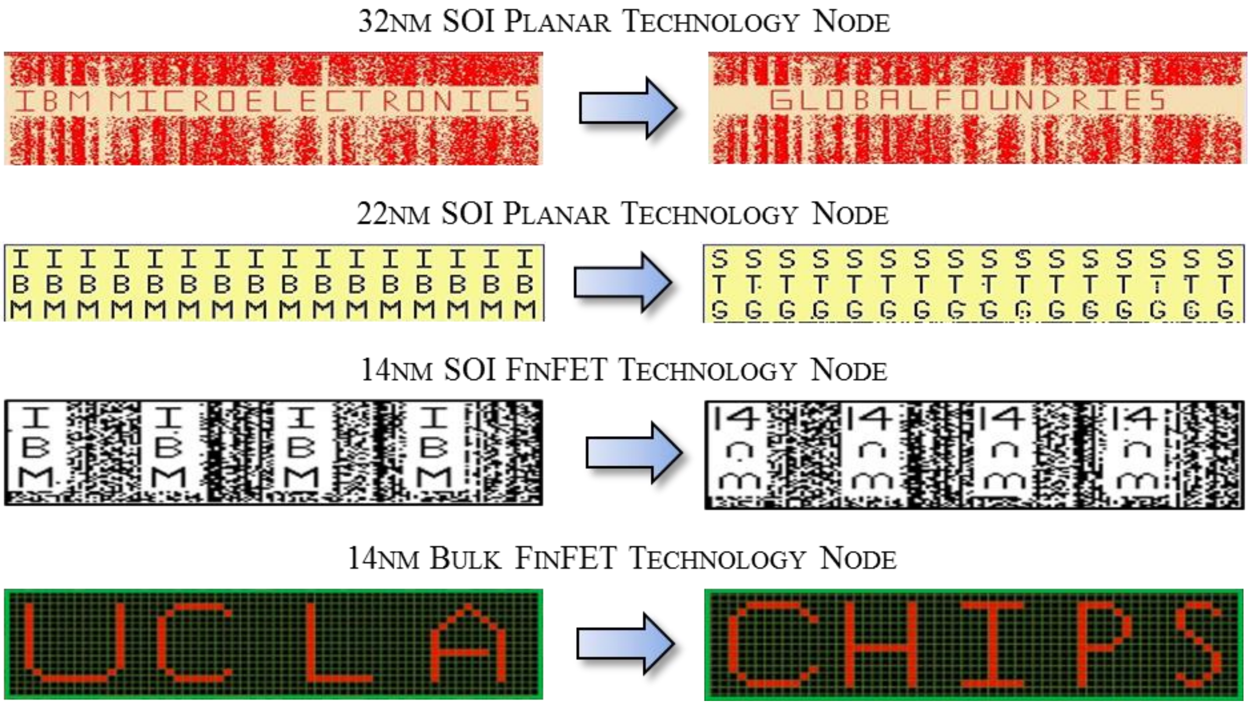

32nm SOI Planar Technology Node
IBM MICROELECTRONICS
GLOBALFOUNDRIES
22nm SOI Planar Technology Node
14nm SOI FinFET Technology Node
14nm Bulk FinFET Technology Node
UCLA
CHIPS

Fig. 4.15. Fully functional CTT eNVM arrays integrated in 32 nm SOI, 22 nm SOI, 14 nm SOI, and 14 nm bulk technology platforms: Bit patterns are written followed by an erase and re-write of alternate bit patterns.

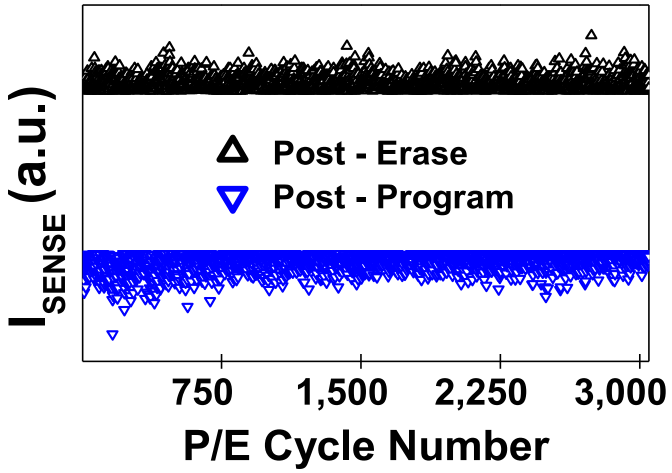


Fig. 4.16. P/E cycling of 7 nm FinFET CTTs using STAR.

4.5 SUMMARY

In this chapter, the fundamental understanding and technological breakthroughs required for employing CTTs as a multi-time programmable (MTP) embedded non-volatile memory (eNVM) for advanced HKMG CMOS technology nodes are outlined. The “initialization” technique, which helps avoid over-programming and consequently reduce memory window drift, is introduced. An erase technique, called “Self-heating Temperature Assisted eRase” (STAR), is introduced which enables 100% erase efficiency, using lower voltage and shorter time, in turn significantly enhancing the P/E endurance of CTTs. For the first time, an endurance of > 10^4 P/E cycles has been demonstrated using CTTs in 14 nm FinFET technology. Data retention lifetime of > 10 years at 125 °C and scalability to 7 nm have been confirmed: 100% erase efficiency and P/E cycling of 7 nm FinFET CTTs using STAR are shown in Fig. 4.16. Circuit design aspects, including sensing techniques, are discussed in [13], [14], [15].

REFERENCES

[1] S. Narasimha, P. Chang, C. Ortolland, D. Fried, E. Engbrecht, K. Nummy, P. Parries, T. Ando, M. Aquilino, N. Arnold, R. Bolam, J. Cai, M. Chudzik, B. Cipriany, G. Costrini, M. Dai, J. Dechene, C. DeWan, B. Engel, M. Gribelyuk, D. Guo, G. Han, N. Habib, J. Holt, D. Ioannou, B. Jagannathan, D. Jaeger, J. Johnson, W. Kong, J. Koshy, R. Krishnan, A. Kumar, M. Kumar, J. Lee, X. Li, C-H. Lin, B. Linder, S. Lucarini, N. Lustig, P. McLaughlin, K. Onishi, V. Ontalus, R. Robison, C. Sheraw, M. Stoker, A. Thomas, G. Wang, R. Wise, L. Zhuang, G. Freeman, J. Gill, E. Maciejewski, R. Malik, J. Norum and P. Agnello, "22nm High-Performance SOI Technology Featuring Dual-Embedded Stressors, Epi-Plate High-K Deep-Trench Embedded DRAM and Self-Aligned Via 15LM BEOL," *IEEE IEDM*, 2012, pp. 3.3.1–3.3.4.

[2] F. Khan, E. Cartier, C. Kothandaraman, J. C. Scott, J. Woo and S. S. Iyer, "The Impact of Self-Heating on Charge Trapping in High-*k*-Metal-Gate nFETs," *IEEE Electron Device Letters*, vol. 37, no. 1, pp. 88-91, 2016.

[3] E. Cartier, B. P. Linder, V. Narayanan and V. K. Paruchuri, "Fundamental understanding and optimization of PBTI in nFETs with SiO_2/HfO_2 gate stack," *IEEE IEDM*, 2006, pp. 1-4.

[4] A. Kerber, S. Krishnan and E. Cartier, "Voltage Ramp Stress for Bias Temperature Instability Testing of Metal-Gate/High-*k* Stacks," *IEEE Electron Device Letters*, vol. 30, no.12, pp.1347-1349, 2009.

[5] E. Cartier and A. Kerber, "Stress-Induced Leakage Current and Defect Generation in nFETs with HfO_2/TiN Gate Stacks during Positive-Bias Temperature Stress", *IEEE IRPS*, 2009, pp. 486-492.

[6] F. Crupi, R. Degraeve, A. Kerber, D.H. Kwak and G. Groeseneken, “Correlation between Stress-Induced Leakage Current (SILC) and the HfO_2 Bulk Trap Density in a SiO_2 / HfO_2 Stack,” *IEEE IRPS*, 2004, pp. 181-187.

[7] F. Khan, M. S. Han, D. Moy, R. Katz, L. Jiang, E. Banghart, N. Robson,T. Kirihata, J. C. S. Woo and S. S. Iyer, “Design Optimization and Modeling of Charge Trap Transistors (CTTs) in 14 nm FinFET Technologies,” *IEEE Electron Device Letters*, vol. 40, no. 7, pp. 1100-1103, 2019.

[8] E. P. Gusev and C. P. D’Emic, “Charge detrapping in HfO_2 high-κ gate dielectric stacks,” *Applied Physics Letters*, vol. 83, no.25, pp.5223-5225, 2003.

[9] F. Khan *et al*., “Program and Erase Memory Structures”, U.S. Patent app. # 16/047,529.

[10] F. Khan *et al*., “Charge Trap Memory Devices”, U.S. Patent app. # 16/781,527.

[11] F. Khan, D. Moy, D. Anand, E. H.-Schroeder, R. Katz, L. Jiang, E. Banghart, N. Robson and T. Kirihata, “Turning Logic Transistors into Secure, Multi-Time Programmable, Embedded Non-Volatile Memory Elements for 14 nm FINFET Technologies and Beyond,” *IEEE Symposium on VLSI Technology*, Kyoto, Japan, 2019, pp. T116-T117.

[12] F. Khan, E. Hunt-Schroeder, D. Moy, D. Anand, R. Katz, D. Leu, J. Fifield, N. Robson, S. Ventrone and T. Kirihata, “A Multi-Time Programmable Embedded Memory Technology in a Native 14nm FINFET Process using Charge Trap Transistors (CTTs),” *Proceedings of the Government Microcircuit Applications & Critical Technology (GOMACTech) Conference*, March 2019.

[13] E. Hunt-Schroeder, D. Anand, J. Fifield, M. Roberge, D. Pontius, M. Jacunski, K. Batson, M. Deming, F. Khan, D. Moy, A. Cestero, R. Katz, Z. Chbili, E. Banghart, L. Jiang, B. Jayaraman, R. R. Tummuru, R. Raghavan, A. Mishra, N. Robson and T. Kirihata, "14nm FinFET 1.5Mb Embedded High-K Charge Trap Transistor One Time Programmable Memory Using Differential Current Sensing," *IEEE Solid-State Circuits Letters*, vol. 1, no. 1, March 2019.

[14] Balaji Jayaraman, Derek Leu, Janakiraman Viraraghavan, Alberto Cestero, Ming Yin, John Golz, Rajesh Reddy Tummuru, Ramesh Raghavan, Dan Moy, Thejas Kempanna, Faraz Khan, Toshiaki Kirihata and Subramanian S. Iyer, "80-kb Logic Embedded High-K Charge Trap Transistor-Based Multi-Time-Programmable Memory With No Added Process Complexity," *IEEE Journal of Solid-State Circuits*, vol. 53, no. 3, pp. 949-960, March 2018.

[15] 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 and 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," *IEEE Symp. VLSI Circuits*, 2016, pp. 1-2.

5. MODELING AND RELIABILITY CONSIDERATIONS

In this chapter, introduced is a compact model that accurately captures the charge trapping behavior in CTTs. The said model explicitly describes the dependence of the resultant device threshold voltage shifts (ΔV_T) on programming time, the vertical electric field, as well as the self-heating temperature during the programming operation and is demonstrated to have excellent agreement with a wide range of experimental results; experiments are performed on hardware in a 14 nm FinFET technology platform [1]. Nominal nFET devices with a gate length of 14 nm and EOT of ~1.3 nm are used.

A key consideration for the development and adaptation of any technology, in additional to performance, is its reliability. In addition to optimizing design and operation conditions to meet the specifications and requirements of any given target application of CTT eNVM, there are certain reliability metrics that one must be cognizant of in terms of device degradation and breakdown. Gate leakage current, dielectric breakdown, and electromigration are such issues that are sometimes encountered in CTT eNVM. The specific concerns and techniques to alleviate them are also discussed in this chapter.

5.1 MODELING THE CTT CHARACTERISTICS

Charge trapping in HfO_2 gate dielectric has been studied extensively since the advent of HKMG devices. V_T shifts that occur due to charge trapping under positive gate bias are referred as “Positive Bias Temperature Instability (PBTI)”. Models that fairly accurately capture PBTI behavior have been developed over the years [2], [3], [4]. However, the aforementioned models do not explicitly capture the effect of self-heating enhanced charge trapping, which is significantly different from the so called PBTI charge trapping as discussed in detail and demonstrated by experimental data in previous chapters. In this work, a comprehensive compact model for self-heating enhanced charge trapping, using the fundamental framework of the existing models has been developed. The said model is shown to describe and predict the CTT programming behavior very accurately.

Time dependence of the charge trapping has been modeled by a power law [2]. A more generalized model for ΔV_T which allows for characterizing the extrapolated maximum possible ΔV_T, 'A', and the characteristic time constant, τ, of the temporal evolution of the device $V_{T,}$ is given by the following expression:

$$\Delta V_T = A.(1 - e^{-(t/\tau_0)^{\beta}}) \tag{5.1}$$

This model assumes a continuous distribution in τ, a function of the capture cross section, where τ_0 is related to the peak in the τ distribution and β is a measure of the width of the distribution: The value of β approaches unity as the distribution width decreases i.e. β=1 implies that the capture cross section has discrete values with no distribution. Additionally, as can be observed from (5.1), the value of 'A' gives the saturation level of the ΔV_T, the experimentally achievable maximum value of which is of course limited by physical limitations such as dielectric breakdown.

β is found to have values between ~0.25 and ~0.5 with programming in the absence of self-heating yielding the lowest values and higher temperatures resulting in slightly higher values. τ_0 is found to decrease logarithmically with programming temperature with values ranging between ~10 s (for room temperature programming) and ~20 ms (for high temperature programming). ΔV_T vs. t_P measured from several different bitcell designs and many different programming conditions is shown in Fig. 5.1, with the values of ΔV_T calculated from the model given by (5.1) overlaid; the model shows excellent agreement with experimental data for a wide range of programming conditions (essentially covering all practical operation conditions for CTT eNVM) and across all the different bitcell designs.

The coefficient 'A' is a function of temperature (determined by the product of R_{th} and the power, $I_{ch}\times V_D$) as well as the electric field (V_G). In order to decouple the impact of self-heating temperature from the effect of electric field, CTT bitcells with various different layouts (as discussed in detail in Chapter 3), and in turn different R_{th} values, as discussed in the previous section, are characterized in detail. In other words, differences in the behaviors of different bitcells under identical programming conditions can be attributed to the differences in their R_{th}. It is found that the voltage acceleration of charge trapping (ΔV_T) is accurately described by a power law. An exponential relationship has been used to model the charge trapping behavior before, but such dependence does not describe the behavior well over a wide voltage range [2]. The temperature acceleration, however, is found to be accurately described by an exponential temperature dependence. 'A' can therefore be expressed as follows:

$$A = d.e^{gT}.V_G^m \tag{5.2}$$

The value of m, which is gate stack dependent (determined by parameters such as interfacial layer and high-k dielectric thickness) is found to be ~7. This is consistent with the reported values in previous literatures. The temperature coefficient, g, is determined to be $\sim 2 \times 10^{-2}$. The coefficient d is determined to be to the order of 10^{-7}, which is expected and consistent with hardware results showing very small ΔV_T values in the absence of self-heating or for small values of V_G. The temperature and V_G dependencies of 'A' (i.e. 'A' normalized by the V_G dependence and temperature dependence, respectively), extracted from experimental results from devices with various different layouts programmed up to the target ΔV_T using many different programming conditions, are shown in Fig. 5.2 (a) and Fig. 5.2 (b), respectively. Overlaid on the same graphs are the corresponding values of normalized ‘*A*’ as predicted by the model given by (5.2); the model shows excellent agreement with hardware data.

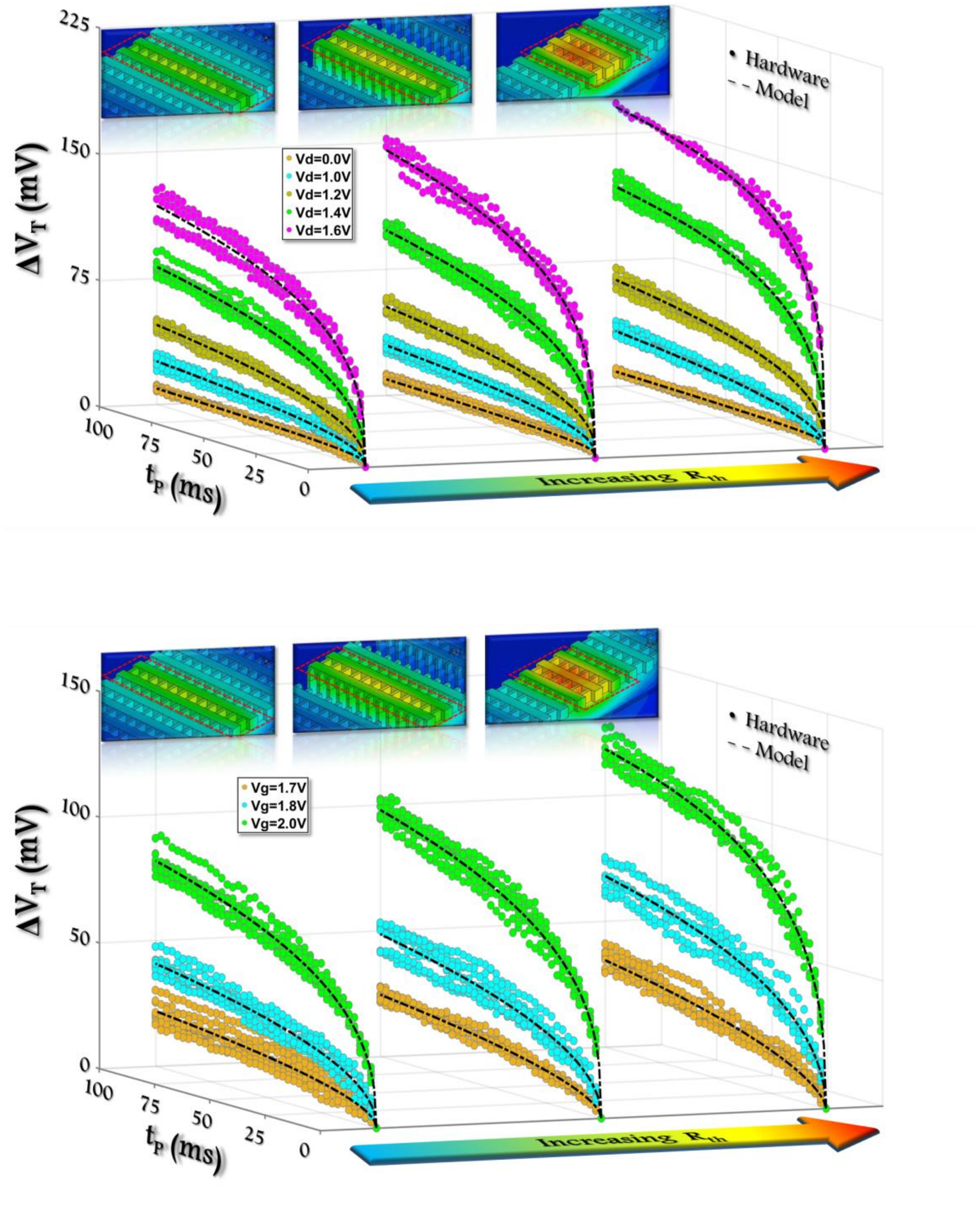


Fig. 5.1. ΔV_T vs. t_P measured from different bitcell designs (shown above their respective datasets) programmed using (a) various V_D, V_G=2V and (b) various V_G, V_D=1.4V (hardware data: colored dots, model: black dashed lines).

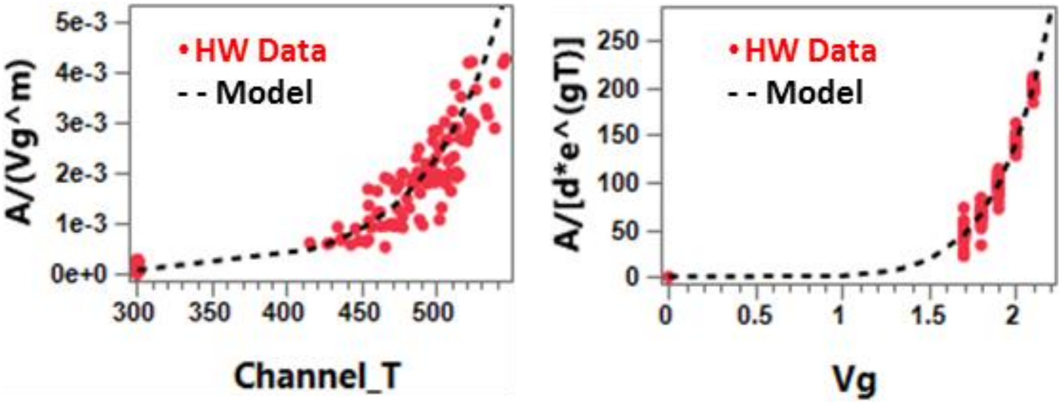


Fig. 5.2. (a) Temperature dependence of '*A*' and (b) V_G dependence of '*A*'.

5.2 RELIABILITY CONSIDERATIONS

Understanding and addressing device reliability is of great importance in any technology. While reliability can be significantly improved by optimizing operation conditions and bitcell layout in the CTT eNVM technology, as discussed in detail in previous chapters, issues such as gate leakage current, dielectric breakdown, and electromigration are nonetheless sometimes encountered. These concerns, and techniques to effectively address them, are discussed here.

It has been observed that gate leakage current in CTTs increases with increasing threshold voltage shift (ΔV_T). This is expected, as an increase in trap density in the HfO_2 layer caused by the stress during the P/E operations results in an increase in SILC (stress-induced leakage current) and is explained by trap-assisted tunneling (TAT) through the defects [5], [6], [7]. Fig. 5.3 shows the increase in the off-state gate leakage current ($I_{G\text{-}OFF}$) measured as a function of ΔV_T. A similar trend is seen for the on-state gate leakage current ($I_{G\text{-}ON}$). It is clear that there is a trade-off between the memory window (ΔV_T) and the gate leakage current. $I_{G\text{-}OFF}$ is defined as the gate leakage current when a particular cell in a memory array has a low V_G and a high V_D (a biasing condition that a cell is subjected to when it is not being read but shares the bitline with another cell that is being read). $I_{G\text{-}ON}$ is defined as the gate leakage current during the read operation of a cell.

Additionally, since the highest vertical field across the gate dielectric during the programming operation is at the source side (V_G=high, Vs=0V, V_D=high), as expected the increase in gate leakage is higher on the source side, as compared to the drain side of the device. The preceding has been confirmed with reverse- vs. forward-mode reads: a forward-mode read (where V_{GS}=high and V_{GD}=low) results in a higher $I_{G\text{-}ON}$ as compared to a reverse-mode read (where V_{GD}=high and V_{GS}=low). The opposite is of course true for $I_{G\text{-}OFF}$, which is higher in a reverse-mode read as compared to a forward-mode read. Since the sum of $I_{G\text{-}OFF}$ from all devices sharing the same bitline in an array impacts the total signal-to-noise-ratio, $I_{G\text{-}OFF}$ can limit the number of wordlines per bitline and in turn the array bit density. Therefore, read conditions favoring lower $I_{G\text{-}OFF}$ are generally preferred.

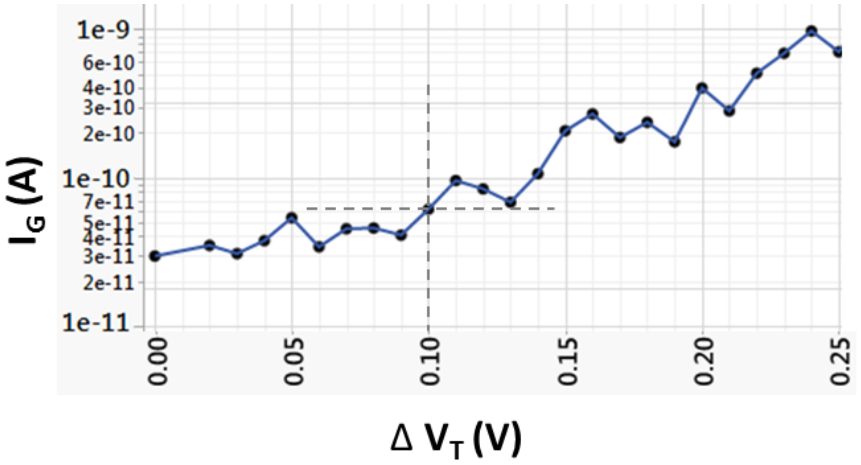


Fig. 5.3. Off-state gate leakage current measured as a function of the CTT threshold voltage shift (ΔV_T).

While $I_{G\text{-}OFF}$ and $I_{G\text{-}ON}$ at any read condition are several orders of magnitude smaller than the channel current, the maximum target ΔV_T nonetheless needs to be considered. The target ΔV_T, which ultimately depends on the particular application and corresponding memory window and data retention specifications, is typically ~100-120mV for most digital applications requiring 10 year data retention at 125 °C. Similarly, the optimized read conditions will depend on the particular array design, sense amplifier design, and application.

The SILC level and the SILC generation rate strongly increase with the applied voltage [6]. In addition, the total time under high bias stress is also a factor that contributes to SILC and gate dielectric breakdown. By optimizing the CTT bitcell design and operation conditions and implementing innovative techniques such as STAR [8], as demonstrated and discussed in detail in Chapters 3 and 4, the required biases, currents, and times for P/E operations can be reduced, which in turn significantly alleviates device reliability and breakdown concerns and drastically improves the P/E endurance of the CTT MTPM.

Relatively high levels of current through the device, flowing in the same direction, during the programming and erase (using the STAR technique) operations can cause electromigration in some CTT bitcells, particularly the source contact metal. Additionally, a relatively high field on one side of the device during the program as well as the erase operations causes time-dependent dielectric breakdown (TDDB) concerns. These issues can cause random bit failures in CTT memory arrays and must be addressed in order to improve the reliability of the CTT eNVM technology and reduce the number of ECC bits required.

A high channel current during the CTT programming operation is required to enable self-heating enhanced charge trapping. At the same time, a high current during the erase operation (i.e. the STAR technique) is absolutely necessary for the CTT MTPM; an endurance of > 10^4 P/E cycles has been demonstrated using the STAR technique (as compared to an endurance of < 15 P/E cycles using the conventional erase method). To further improve the endurance of the CTT eNVM as well as reduce the random bit failure rate, the AC-STAR (“Alternating Current Self-heating Temperature Assisted eRase”) [9] technique has been developed.

Like in the STAR approach, AC-STAR employs self-heating in the device by utilizing the source-substrate-drain structure of the CTT memory cell as a parasitic BJT to pass a short current pulse through the body of the device during the erase operation, while simultaneously enabling a negative gate-to-substrate bias. However, the AC-STAR approach alternates the bias direction between the source and the drain i.e. the emitter and collector terminals of the parasitic BJT are alternated. This technique has three major advantages: *(i)* electromigration is mitigated, reducing the random bit failure rate significantly, *(ii)* the erase efficiency and reliability is improved due to a more uniform erase, and *(iii)* the risk of breakdown due to the gate-to-drain bias is reduced as the high bias is now shared between gate-drain and gate-source due to the alternating bias. Moreover, gate dielectric breakdown probability (caused by TDDB) is reduced due to reduced high voltage stress time. Other intrinsic parameters such as gate leakage currents are also benefitted.

Implementation of the AC STAR technique almost completely eliminates the electromigration issue. A TEM cross section of a 14 nm FinFET CTT that has broken down due to electromigration of the source contact metal (W) at some point during the P/E cycling is shown in Fig. 5.4(a). A TEM cross section of a 14 nm FinFET CTT, cycled 10^4 times using the AC-STAR technique, demonstrating no signs of electromigration is shown in Fig. 5.4(b). Indeed, the random bit failure rate and correspondingly the CTT memory array yield are observed to have been significantly improved by implementation of the AC STAR technique.

Furthermore, like in the AC-STAR technique, the high bias node can also be alternated between the source and the drain of the CTT during the programming operation. This technique, called "AC Programming" or "AC-PRG", further alleviates gate dielectric breakdown concerns by alternating the high potential difference between gate-drain and gate-source during the programming operation.

Further details on the STAR, AC-STAR, and AC-PRG techniques, including circuit implementations, can be found in [8] and [9].

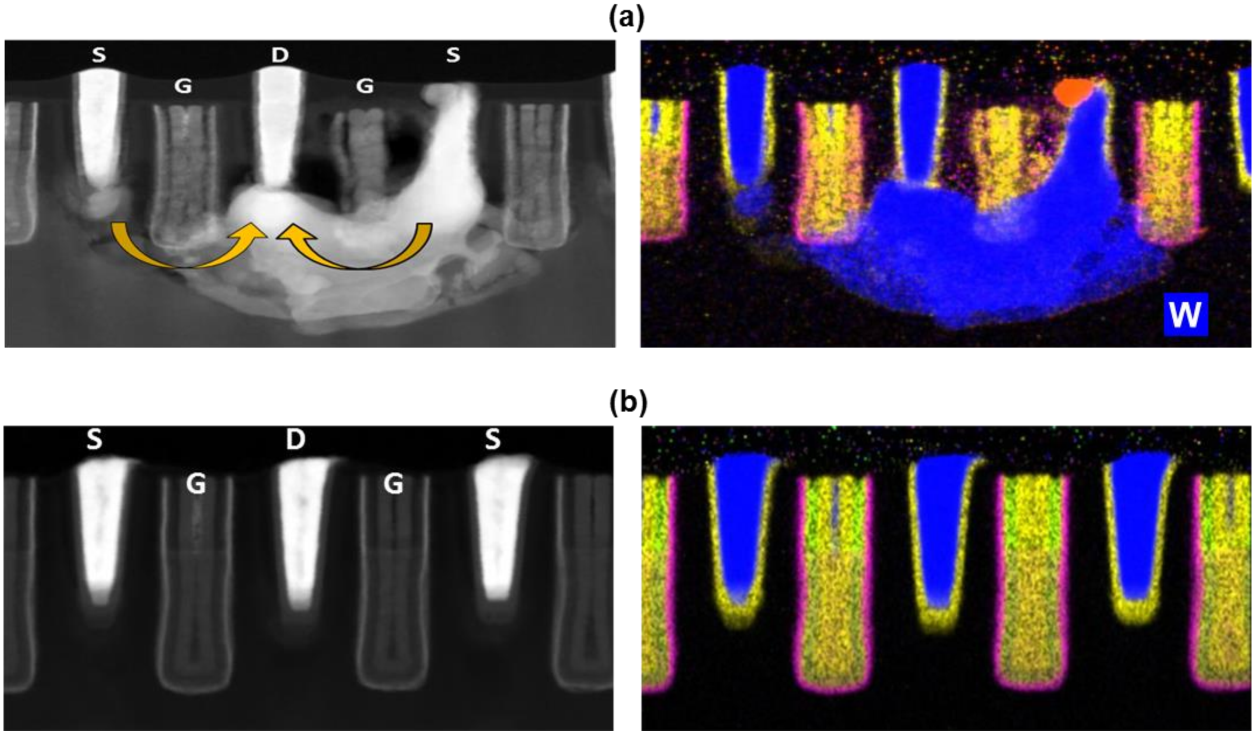


Fig. 5.4. (a) TEM cross-section of a 14 nm FinFET CTT showing electromigration of the source contact metal (W). The electron current flows during the STAR operation are depicted by arrows. (b) A TEM cross section of a 14 nm FinFET CTT, cycled 10^4 times using the AC-STAR technique, demonstrating no signs of electromigration. The bright regions in the left-side images are W. Elemental analyses of the respective cross-sections are shown in the right-side images wherein the W is displayed in blue color.

5.3 SUMMARY

A compact model that can be used to accurately characterize and predict the charge trapping behavior in CTTs is introduced in this chapter. The model explicitly describes and decouples the electric-field and self-heating temperature dependencies of charge trapping in CTTs, which is also applicable to charge trapping in HKMG devices in general. Such a compact model is useful for optimization of operation conditions as well as bitcell design of the CTT eNVM. Excellent agreement between the model and experimental data from several different bitcell designs and a wide range of programming conditions, covering all practical operation conditions for CTT eNVM, has been demonstrated.

Also included in this chapter is a discussion on reliability concerns such as gate leakage current, dielectric breakdown, and electromigration in the CTT eNVM technology. Innovations, such as STAR, AC-STAR, and AC-PRG techniques, that significantly alleviate the said concerns are demonstrated.

REFERENCES

[1] J. Singh, A. Bousquet, J. Ciavatti, K. Sundaram, J. S. Wong, K. W. Chew, A. Bandyopadhyay, S. Li, A. Bellaouar, S. M. Pandey, B. Zhu, A. Martin, C. Kyono, J.-S. Goo, H. S. Yang, A. Mehta, X. Zhang, O. Hu, S. Mahajan, E. Geiss, S. Yamaguchi S. Mittal, R. Asra, P. Balasubramaniam, J. Watts, D. Harame, R. M. Todi, S. B. Samavedam and D.K. Sohn, "14nm FinFET Technology for Analog and RF Applications," *IEEE Symp. VLSI Technol.*, 2017, pp. T140-T141.

[2] A. Kerber and E. Cartier, "Reliability Challenges for CMOS Technology Qualifications With Hafnium Oxide/Titatnium Nitride Gate Stacks," *IEEE Transactions on Device and Materials Reliability*, vol. 9, no. 2, pp. 147-162, 2009.

[3] S. Zafar, A. Callegari, E. Gusev and M. V. Fischetti, "Charge Trapping in High *k* Gate Dielectric Stacks," *IEDM Tech. Dig.*, 2002, pp. 517-520.

[4] G. Ribes, J. Mitard, M. Denais, S. Bruyere, F. Monsieur, C. Parthasarathy, E. Vincent and G. Ghibaudo, "Review on High-*k* Dielectrics Reliability Issues," *IEEE Transactions on Device and Materials Reliability*, vol. 5, no.1, pp. 5-19, March 2005.

[5] E. Cartier and A. Kerber, "Stress-Induced Leakage Current and Defect Generation in nFETs with HfO_2/TiN Gate Stacks during Positive-Bias Temperature Stress," *IEEE IRPS*, 2009, pp. 486-492.

[6] F. Crupi, R. Degraeve, A. Kerber, D.H. Kwak and G. Groeseneken, "Correlation between Stress-Induced Leakage Current (SILC) and the HfO_2 Bulk Trap Density in a SiO_2 / HfO_2 Stack," *IEEE IRPS*, 2004.

[7] R. O’Connor, L. Pantisano, R. Degraeve, T. Kauerauf, B. Kaczer, Ph. J. Roussel and G. Groeseneken, “SILC Defect Generation Spectroscopy in HfSiON Using Constant Voltage Stress and Substrate Hot Electron Injection,” *IEEE International Reliability Physics Symposium*, 2008, pp. 324-329.

[8] F. Khan *et al*., “Program and Erase Memory Structures”, U.S. Patent app. # 16/047,529.

[9] F. Khan *et al*., “Charge Trap Memory Devices”, U.S. Patent app. # 16/781,527.

6. SUMMARY, CONCLUSIONS, AND OUTLOOK

The Charge Trap Transistor (CTT) is a novel embedded non-volatile (eNVM) technology that turns as-fabricated standard logic transistors into multi-time programmable memory elements for advanced high-*k*/metal (HKMG) CMOS technology nodes, without the need for any additional processes or masks. The fundamental device physics and principles of operation of CTTs and viability of the CTT eNVM technology for commercial applications have been demonstrated in this work. By implementing the fundamental understanding, principles of operation, and innovations presented in this work, a commercially available CTT eNVM product that is capable of operating at military grade temperatures has already been deployed. Implementation of CTT eNVM has been demonstrated in 32 nm SOI planar, 22 nm SOI planar, 14 nm SOI FinFET, and 14 nm bulk FinFET production technologies. Scalability of the CTT technology to 7 nm nodes has also been demonstrated.

The CTT technology employs as-fabricated standard logic transistors as eNVM elements by exploiting the phenomena of self-heating enhanced charge trapping. The fundamentals of self-heating enhanced charge trapping in HKMG CMOS transistors and the corresponding implications for memory applications have been analyzed in detail. It is demonstrated that the magnitude as well as the stability (retention) of the trapped charge is significantly enhanced with device self-heating, making the resultant device threshold voltage shifts (ΔV_T) large and stable enough for non-volatile memory applications requiring high-temperature operation: data retention lifetime of > 10 years at 125 °C has been demonstrated. Also presented in this work are techniques to optimize the CTT bitcell design for enhancing the programming efficiency. In particular, how device layout can be manipulated to maximize self-heating assisted charge trapping is discussed.

Furthermore, the fundamental understanding and technological breakthroughs required for employing CTTs as a multi-time programmable memory (MTPM) for advanced HKMG CMOS technologies are presented. The “initialization” technique, which helps avoid over-programming and consequently reduce memory window drift, is introduced. An erase technique, called “Self-heating Temperature Assisted eRase” (STAR), is introduced which enables 100% erase efficiency, using lower voltage and shorter time, in turn significantly enhancing the P/E endurance of the CTT eNVM.

Also included this work is a compact model that accurately characterizes and predicts the charge trapping behavior in CTTs. Additionally, reliability concerns in the CTT eNVM technology and techniques to effectively address those concerns have been discussed.

Potential applications of the CTT technology include hardware security, reconfigurable on-chip encryption key storage, firmware, BIOS, chip ID, configuration memory, redundancy, repair, performance tailoring, and field configurability. Moreover, the CTT array in its native (unprogrammed) state measures very well as an entropy source for potential PUF (Physically Unclonable Function) applications. As demonstrated in Fig. 6.1, the hamming weight, intra-instance hamming distance, and inter-instance hamming distance are all very close to an ideal entropy source. Implementation of CTT arrays as PUFs for authentication, identification, anti-counterfeiting, secure boot, and cryptographic IP is another area of strong interest and active investigation.

In addition to the numerous digital applications, CTTs can also be utilized as an analog memory for applications like neuromorphic computing for machine learning (ML) and artificial intelligence (AI). CTTs demonstrate excellent analog memory characteristics: the ΔV_T can be modulated, with a high resolution, back and forth within the memory window (Fig. 6.2). Indeed, researchers have already proposed several viable applications of CTTs as an analog memory [1], [2], [3], [4]. A CTT neural network based analog inference engine is expected to be significantly more energy efficient than digital inference engines and with similar performance. The inference accuracy of a CTT based analog inference engine has been shown to be significantly better than any other analog inference engine (using PCM, Memristor, RRAM, etc.) published thus far [2].

Additionally, being a three-terminal device with a high subthreshold slope i.e. a large change in output current for a small input voltage change, operation in the subthreshold region results in a large signal ON/OFF ratio which makes the CTT suitable for general purpose subthreshold logic applications.

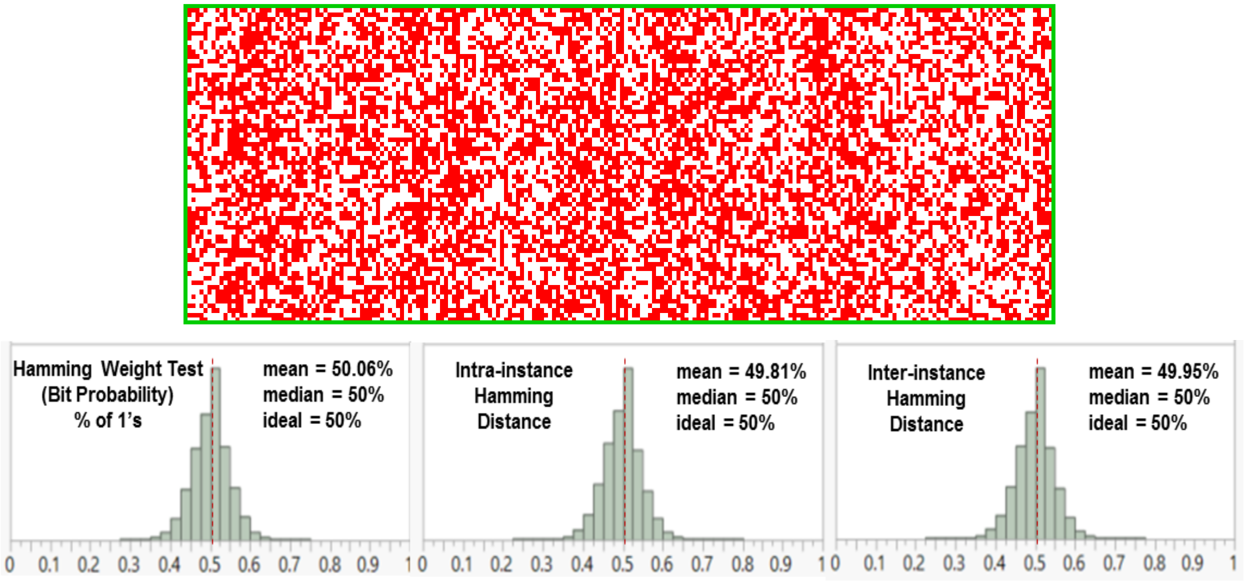


Fig. 6.1. (Top panel) A 14 nm FinFET CTT array in its native state. (Bottom panel) Hamming weight, intra-instance hamming distance, and inter-instance hamming distance all show a nearly ideal entropy source. Tests were performed on 354 Mb of data from 59 different chips.

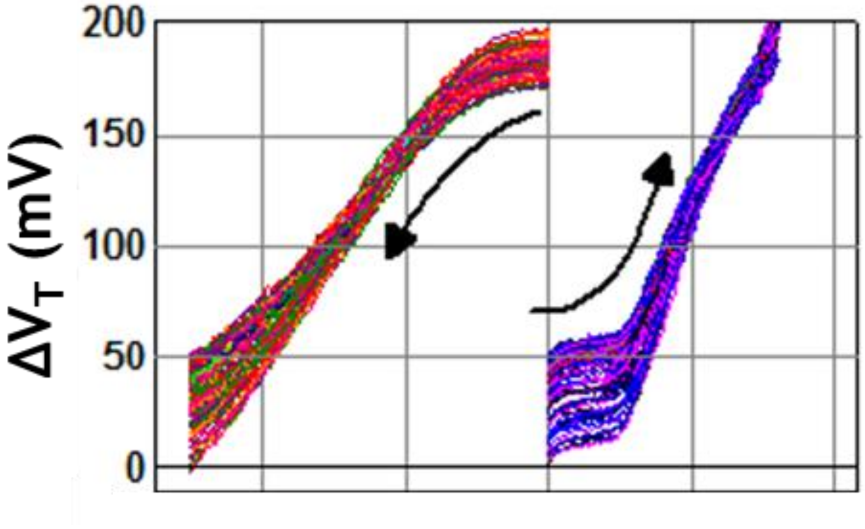


Fig. 6.2. Demonstration of the CTT as an analog memory. ΔV_T is modulated back and forth, 1000 times, between 0 and ~200mV in ~2mV increments. Each small dot represents a unique ΔV_T level and each color represents a different program or erase cycle.

In summary, the CTT technology is a 100% logic compatible eNVM solution (plug-in technology) that offers the following demonstrated advantages and features:

- No additional processing or masks needed
- Re-programmable: > 10^4 P/E cycle endurance demonstrated
- Logic compatible voltage operation: ~2V max
- Scalable: feasibility down to the 7 nm node demonstrated
- High operation temperature range: -55 to 125 °C
- Robust data-retention: 10 years @ 125 °C
- Secure
- Digital as well as analog memory applications

Extension of the CTT eNVM endurance beyond 10^4 P/E cycles and further improvement of the P/E efficiency i.e. lowering the required power and/or time for each P/E operation will require further innovation and are subjects that warrant further research work.

References

[1] Y. Du *et al*., “An Analog Neural Network Computing Engine Using CMOS-Compatible Charge-Trap-Transistor (CTT),” *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 38, no. 10, Oct. 2019, pp. 1811-1819.

[2] X. Gu, Z. Wan and S. S. Iyer, “Charge-Trap Transistors for CMOS-Only Analog Memory,” *IEEE Transactions on Electron Devices*, vol. 66, no. 10, pp. 4183-4187, Oct. 2019.

[3] S. Ma, M. Donato, S. K. Lee, D. Brooks and G.-Y. Wei, “Fully-CMOS Multi-Level Embedded Non-Volatile Memory Devices With Reliable Long-Term Retention for Efficient Storage of Neural Network Weights,” *IEEE Electron Device Letters*, vol. 40, no. 9, pp. 1403-1406, Sep. 2019.

[4] M. Donato, B. Reagen, L. Pentecost, U. Gupta, D. Brooks and G. Wei, “On-chip Deep Neural Network Storage with Multi-level eNVM,” *55th ACM/ESDA/IEEE Design Automation Conference (DAC),* 2018, pp. 1-6.