Title

Design Optimization and Modeling of Charge Trap Transistors (CTTs) in 14 nm FinFET Technologies

Author
Khan, Faraz

Publication Date
2019-05-01

Design Optimization and Modeling of Charge Trap Transistors (CTTs) in 14 nm FinFET Technologies

Faraz Khan, Min Soo Han, Dan Moy, Robert Katz, Liu Jiang, Edmund Banghart, Norman Robson, Toshiaki Kirihata, Jason C. S. Woo, *Fellow, IEEE*, and Subramanian S. Iyer, *Fellow, IEEE*

***Abstract*—The Charge Trap Transistor (CTT) technology is an emerging memory solution that turns as-fabricated high-*k*/metal gate (HKMG) logic transistors into secure, embedded non-volatile memory (eNVM) elements with excellent data retention and operation capability at military grade temperatures. In other words, the CTTs offer a completely process-free and mask-free eNVM solution for advanced HKMG CMOS technology nodes. In this letter, bitcell design to enhance programming efficiency and modeling of the charge trapping behavior of CTTs in 14 nm FinFET technology is discussed.**

***Index Terms*—Charge Trap Transistor (CTT), embedded non-volatile memory (eNVM), process-free, mask-free, high-*k*/metal gate (HKMG), CMOS.**

I. Introduction

WHILE need for on-chip non-volatile memory in VLSI technologies continues to grow rapidly, the options have been limited due to integration and scaling challenges as well as operational voltage incompatibilities. eFUSE and anti-fuse [1] technologies require high voltages which are logic incompatible and also face scaling challenges. Other memory technologies like MONOS (metal-oxide-nitride-oxide-silicon) [2] and MRAM (magnetoresistive random access memory) [3] require additional complex processes and masks. Unlike the aforementioned memories, CTTs offer an embedded non-volatile memory (eNVM) solution that requires absolutely no additional processes or masks, operates at logic-compatible voltages (~2V maximum), and is scalable. Chip configuration, repair at wafer and module test and in the field, firmware, and performance tailoring are some applications of CTT eNVM. CTTs also find their applications in security enhancements such as chip ID, authentication, and encryption key storage.

Programming of CTTs is achieved using the fundamental operation principle of 'device self-heating enhanced charge trapping' in standard as fabricated high-*k*/metal gate (HKMG) logic devices [4], [5]; the device threshold voltage (V_T) is modulated by the 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. CTTs are typically programmed using short gate bias (V_G) pulses of 1.8-2.0V with a drain bias (V_D) of 1.4-1.6V, while the source bias (V_S) and the substrate bias (V_X) are at 0V. Previously, it was demonstrated how layout-dependent effects (LDE) in planar devices can be manipulated to modulate and enhance the self-heating effect and in turn the programming efficiency in CTTs [4]. It was shown that 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 narrow channels in parallel.

In this work, for the first time, we demonstrate how the CTT bitcell design can be manipulated to exploit LDE, to significantly enhance the programming efficiency in FinFET-based CTTs; experiments are performed on hardware in a 14 nm FinFET technology platform [6]. Nominal nFET devices with a gate length of 14 nm and EOT of ~1.3 nm are used. Moreover, we introduce a compact model that accurately captures the CTT programming behavior as a function of programming time, the vertical electric field, as well as the self-heating temperature.

II. Bitcell Design for Enhanced Programming Efficiency

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

Manuscript received April 30, 2019; revised May 22, 2019; accepted May 26, 2019. Date of publication May 30, 2019; date of current version June 26, 2019. The review of this letter was arranged by Editor K. J. Kuhn. *(Corresponding author: Faraz Khan.)*

F. Khan is with GlobalFoundries, Advanced Technology Development, East Fishkill, NY 12533 USA, and also with the Electrical and Computer Engineering Department, University of California at Los Angeles, Los Angeles, CA 90095 USA (e-mail: faraz.khan@globalfoundries.com).

M. S. Han, J. C. S. Woo, and S. S. Iyer are with the Electrical and Computer Engineering Department, University of California at Los Angeles, Los Angeles, CA 90095 USA.

D. Moy, R. Katz, L. Jiang, E. Banghart, N. Robson, and T. Kirihata are with GlobalFoundries, Advanced Technology Development, East Fishkill, NY 12533 USA.

Color versions of one or more of the figures in this letter are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LED.2019.2919871

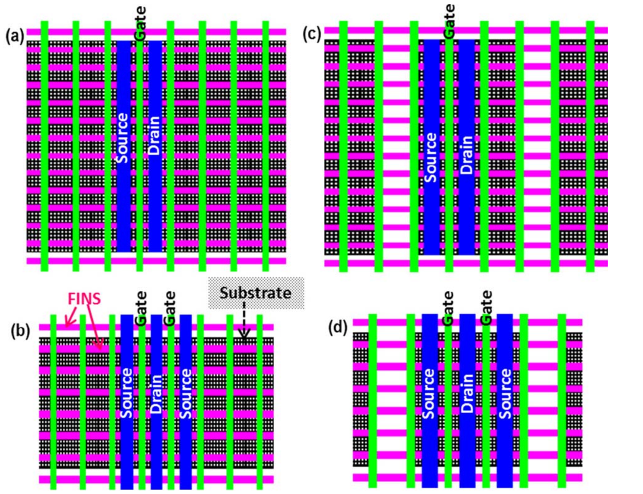

Fig. 1. Top down views of (a) '1 gate × 12 fin', (b) '2 gate × 6 fin', (c) '1 gate × 12 fin' on active "island", and (d) '2 gate × 6 fin' on active "island" CTT bitcell layouts.

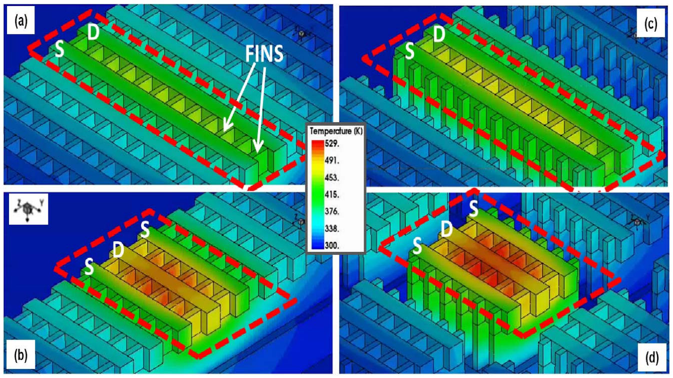

Fig. 2. 3D finite element thermal simulation of the programming operation of CTT bitcell structures with (a) '1 gate × 12 fin', (b) '2 gate × 6 fin', (c) '1 gate × 12 fin' on active "island", and (d) '2 gate × 6 fin' on active "island" layouts.

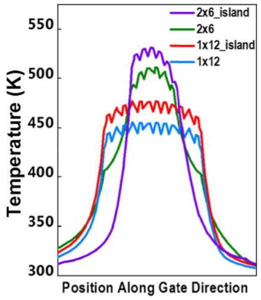

Fig. 3. Thermal profiles, during programming, of the bitcell layouts shown in Fig. 2, along the gate direction.

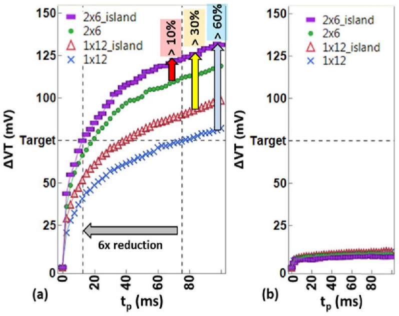

Fig. 4. ΔV_T vs. t_P for CTT bitcells with various layouts. The devices are programmed with (a) $V_G = 2V$, $V_D = 1.4V$, $V_S = 0V$ and (b) $V_G = 2V$, $V_D = V_S = 0V$.

bitcell layouts are fabricated and studied. A '1 gate × 12 fin' bitcell (Fig. 1(a)) and a '2 gate × 6 fin' bitcell (Fig. 1(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. 1 (c) and Fig. 1 (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.

The bitcell steady-state temperatures (achieved within ~40-50 ns) during the programming operation are shown in Fig. 2. The thermal profiles along the gate direction (perpendicular to the fins), at programming conditions, of each of the four bitcells (Fig. 3) 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. 4. 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. 4 (a)). Secondly, in the absence of self-heating (Fig. 4 (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 [4].

III. 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 to as "Positive Bias Temperature Instability (PBTI)". Models that fairly accurately capture PBTI behavior have been developed over the years [7]–[9]. 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 the previous section and demonstrated by the data in Fig. 4. In this work, a comprehensive compact model for self-heating enhanced charge trapping, using the fundamental framework of the

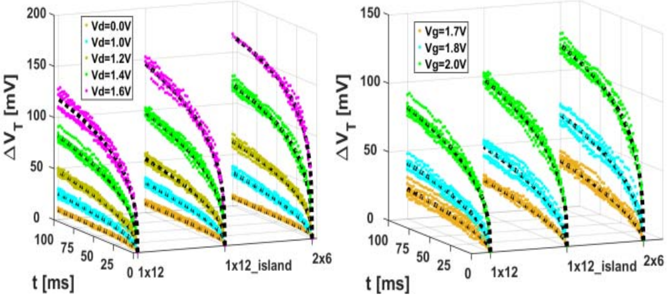

Fig. 5. ΔV_T vs. t_P measured from different bitcell designs programmed using (a) various V_D, $V_G = 2V$ and (b) various V_G, $V_D = 1.4V$ (hardware data: colored dots, model: black dotted lines).

existing models has been developed. The said 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 [7]. 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;

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

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. $\beta = 1$ implies that the capture cross section has discrete values with no distribution. Additionally, as can be observed from (1), the value 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 different bitcell designs and many different programming conditions is shown in Fig. 5, with the values of ΔV_T calculated from the model given by (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, 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 [7]. The temperature acceleration, however, is found to be accurately described by an exponential temperature dependence. 'A' can therefore be expressed as follows

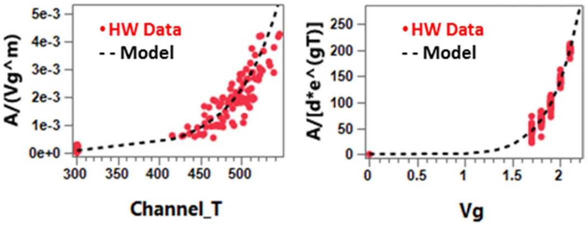

Fig. 6. (a) Temperature dependence of 'A' and (b) V_G dependence of 'A'.

$$A = d \cdot e^{gT} \cdot V_G^m \tag{2}$$

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 upto the target ΔV_T using many different programming conditions, are shown in Fig. 6 (a) and Fig. 6 (b), respectively. Overlaid on the same graphs are the corresponding values of normalized 'A' as predicted by the model given by (2); the model shows excellent agreement with hardware data.

IV. Summary and Conclusions

In this letter, we demonstrate the fundamental understanding and techniques to optimize the CTT eNVM 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 [4], is discussed. Also introduced is a compact model that can be used to accurately characterize and predict the programming behavior of CTTs. 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. A 14 nm FinFET CTT based one-time programmable memory (OTPM) product has already been deployed: 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 [10]. While only programming related aspects of CTTs are discussed in this letter, CTTs can also be employed as multi-time programmable memory (MTPM) elements, which would of course require erasing the programmed devices efficiently. The technological breakthroughs required for implementation of CTTs as an MTPM in 14nm 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 [11].

References

[1] S.-Y. Chou, Y.-S. Chen, J.-H. Chang, Y.-D. Chih, and T.-Y. J. Chang, "A 10 nm 32 Kb low-voltage logic-compatible anti-fuse one-time-programmable memory with anti-tampering sensing scheme," in *Proc. IEEE Int. Solid-State Circuits Conf.*, Feb. 2017, pp. 200–202.

[2] 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 16/14 nm-node embedded flash," in *IEDM Tech. Dig.*, Dec. 2017, pp. 19.3.1–19.3.4.

[3] 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," in *IEEE VLSI Technol. Dig. Tech. Papers*, Jun. 2018, pp. 183–184.

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

[5] F. Khan, E. Cartier, J. Woo, and 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 Lett.*, vol. 38, no. 1, pp. 44–47, Jan. 2017.

[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, "14 nm FinFET technology for analog and RF applications," in *Proc. IEEE Symp. VLSI Technol.*, Jun. 2017, pp. T140–T141.

[7] A. Kerber and E. A. Cartier, "Reliability challenges for CMOS technology qualifications with hafnium oxide/titanium nitride gate stacks," *IEEE Trans. Device Mater. Rel.*, vol. 9, no. 2, pp. 147–162, Jun. 2009.

[8] S. Zafar, A. Callegari, E. Gusev, and M. Fischetti, "Charge trapping in high k gate dielectric stacks," in *IEDM Tech. Dig.*, San Francisco, CA, USA, Dec. 2002, pp. 517–520.

[9] 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 Trans. Device Mater. Rel.*, vol. 5, no. 1, pp. 5–19, Mar. 2005.

[10] 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, "14-nm FinFET 1.5 Mb embedded high-*k* charge trap transistor one time programmable memory using differential current sensing," *IEEE Solid-State Circuits Lett.*, vol. 1, no. 12, pp. 233–236, Dec. 2018.

[11] F. Khan, D. Moy, D. Anand, E. Hunt-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," in *Proc. IEEE Symp. VLSI Technol. Dig. Tech. Papers*, 2019, pp. T116–T117.