UCLA
UCLA Previously Published Works

Title
Charge Trap Transistor (CTT): An Embedded Fully Logic-Compatible Multiple-Time Programmable Non-Volatile Memory Element for High-k-Metal-Gate CMOS Technologies

Permalink
https://escholarship.org/uc/item/1hb0h8bz

Journal
IEEE Electron Device Letters, 38(1)

ISSN
0741-3106 1558-0563

Author
Khan, Faraz

Publication Date
2017

DOI
10.1109/LED.2016.2633490

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Charge Trap Transistor (CTT): An Embedded Fully Logic-Compatible Multiple-Time Programmable Non-Volatile Memory Element for High-*k*-Metal-Gate CMOS Technologies

Faraz Khan, Eduard Cartier, Jason C. S. Woo, *Fellow, IEEE,* and Subramanian S. Iyer, *Fellow, IEEE*

***Abstract*—The availability of on-chip non-volatile memory for advanced high-*k*-metal-gate CMOS technology nodes has been limited due to integration and scaling challenges as well as operational voltage incompatibilities, while its need continues to grow rapidly in modern high-performance systems. By exploiting intrinsic device self-heating enhanced charge trapping in as fabricated high-k-metal-gate logic devices, we introduce a unique multiple-time programmable embedded non-volatile memory element, called the 'charge trap transistor' (CTT), for high-k-metal-gate CMOS technologies. Functionality and feasibility of using CTT memory devices have been demonstrated on 22 nm planar and 14 nm FinFET technology platforms, including fully functional product prototype memory arrays. These transistor memory devices offer high density (~0.144μm^2/bit for 22 nm and ~0.082μm^2/bit for 14 nm technology), logic voltage compatible and low peak power operation (~4mW), and excellent retention for a fully integrated and scalable embedded non-volatile memory without added process complexity or masks.**

***Index Terms*—High-k-metal-gate, CMOS, embedded non-volatile memory.**

I. INTRODUCTION

THERE is an ever-increasing need for on-chip memory in VLSI technologies. In this letter, we demonstrate the application of HfO_2 based high-*k*-metal-gate (HKMG) logic devices, called 'Charge Trap Transistors' (CTTs), as non-volatile memory (NVM) elements for system-on-chip (SoC) applications in state-of-the-art CMOS technologies. Recently, we demonstrated that intrinsic self-heating enhanced charge trapping in HKMG devices can be used to achieve large and stable device threshold voltage (V_T) shifts that are suitable for memory applications [1]. In this work, we demonstrate for the first time that indeed multiple-time programmability is possible for application of CTTs as multiple-time programmable non-volatile memory elements. We discuss the underlying principles of operation and the key factors for optimized operation of the memory devices.

Manuscript received November 16, 2016; revised November 22, 2016; accepted November 25, 2016. Date of publication December 1, 2016; date of current version December 27, 2016. This work was supported in part by DARPA (award FA 8650-16-C-7648), and in part by GlobalFoundries. The review of this letter was arranged by Editor T. Wang.

F. Khan is with the Center for Heterogeneous Integration and Performance Scaling (CHIPS), Electrical Engineering Department, University of California at Los Angeles, Los Angeles, CA 90095 USA, and Global Foundries, Advanced Technology Development, Malta, NY 12020 USA (e-mail: farazkhan@ucla.edu).

E. Cartier is with IBM Research, Thomas J. Watson Research Center, Yorktown Heights, NY 10598 USA.

J. C. S. Woo and S. S. Iyer are with the Center for Heterogeneous Integration and Performance Scaling (CHIPS), Electrical Engineering Department, University of California at Los Angeles, Los Angeles, CA 90095 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.2016.2633490

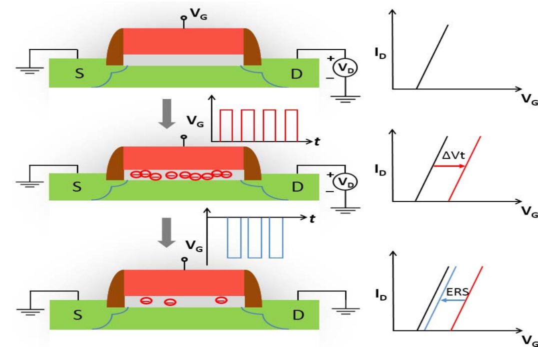

Fig. 1. A schematic depicting the basic operation of a CTT memory device (equally applicable to planar FET as well as FinFET based CTTs).

In addition to being multiple-time programmable (MTP), CTT based memory offers a better alternative to existing one-time programmable (OTP) technologies like eFUSE [2] and gate breakdown anti-fuse [3] as it can be used more effectively for yield improvement, field configurability, performance tailoring, and security enhancements such as chip IDs and on-chip reconfigurable encryption key and firmware storage with lower power, higher density, and higher scalability, at no additional processing cost. This intrinsic self-heating enhanced charge trapping memory solution is applicable to SOI as well as bulk FinFET technologies as self-heating in bulk FinFETs, while generally less than SOI FinFETs, is comparable to SOI planar devices and increases considerably with scaling [4], [5].

II. PRINCIPLES OF OPERATION

A schematic of the basic operation of a CTT memory device is depicted in Fig. 1; The device V_T is modulated by the charge trapped in the high-*k* dielectric of the HKMG device. It must be noted that, while a planar device is used for demonstration, the same principles equally apply to FinFET based CTTs.

In order to understand the dynamic behavior of charge trapping in the devices, we first measured the V_T shifts (ΔV_T) as a function of the programming time (t_p), as seen in Fig. 2(a). We used 1.2um×20nm devices (22nm SOI technology [6]) and programmed them using 2V gate voltage (V_g) pulses while the drain-to-source voltage (V_d) was fixed at 1.3V. ΔV_T vs. stress time with 2V V_g pulses and $V_d = 0$V

0741-3106 © 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

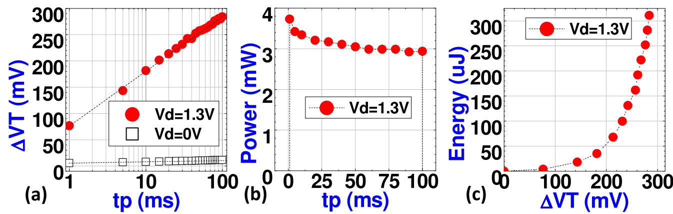

Fig. 2. (a) ΔV_T vs. t_p (V_g = 2V, V_d = 1.3V). ΔV_T for PBTI @ 25°C (V_g = 2V, V_d = 0V) is shown for comparison. (b) Power consumption vs. time during programming. (c) Calculated total energy ($E_p = \int I_d \times V_d \times t_p$) required vs. ΔV_T created.

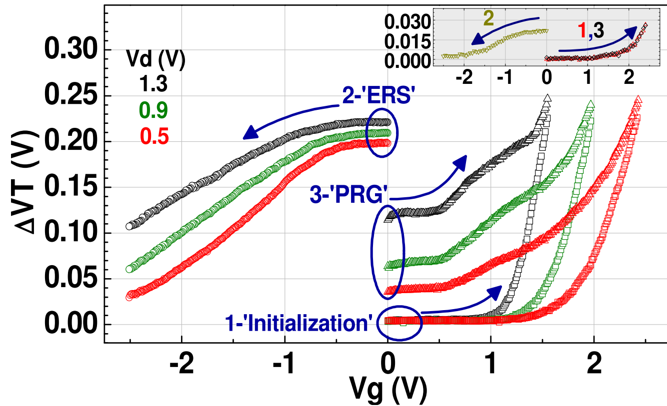

Fig. 3. Measured ΔV_T during 1-'Initialization', 2-'ERS', and 3-'PRG' cycles for various V_d values using PVRS. Inset shows ΔV_T for V_d = 0V PVRS stress (BTI).

($I_{ds} \cong 0$) is also shown for comparison between self-heating enhanced charge trapping at high V_d [1] and conventional Positive Bias Temperature Instability (PBTI) [7]. It is observed that ΔV_T is dramatically enhanced when the transistor is pulsed at high V_d [1] 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. 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 eFUSE elements. Fig. 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. 2(a) i.e. programming efficiency reduces as $t_p$increases.

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 was performed using the Pulsed gate Voltage Ramp Sweep (PVRS) technique (details on PVRS discussed in [1] and [8]), with 10ms V_g pulses of increasing magnitudes in 10mV increments, as demonstrated in Fig. 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 MTP memory 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 [1], [9], and [10], 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 [1], where enhanced charge retention was demonstrated for devices programmed at higher V_d. The slight ΔV_T difference between the end of the 'INIT' 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 [1] and also in section IV. *(iii)* The charge trapping behavior changes after the 'initialization' operation; This is due to the creation of new traps [9], [10], allowing for subsequent programming ('PRG') to the same ΔV_T at lower V_g. This phenomenon has also been reported in [11], 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. 3 inset). It is clearly seen that, without the effect of self-heating, (i)For the same V_g values, 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.

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 [1], 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. Thus, 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 [1].

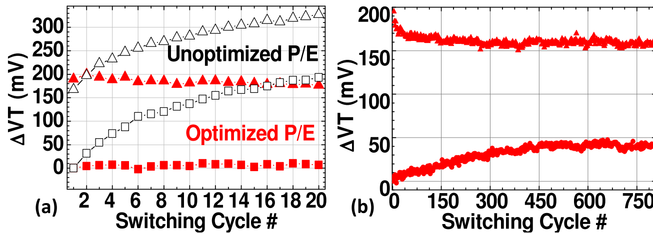

Fig. 4. (a) Memory window vs. switching cycle number comparison between un-optimized P/E (V_{g-INIT} = 2V, V_{d-INIT} = 1.3V, V_{g-PRG} = 2V, V_{d-PRG} = 1.3V, V_{g-ERS} = −2V, open black symbols) and optimized (V_{g-INIT} = 2V, V_{d-INIT} = 1.3V, V_{g-PRG} = 2V, V_{d-PRG} = 1.2V, V_{g-ERS} = −2V, solid red symbols) P/E conditions. (b) 800x P/E cycles using optimized P/E conditions.

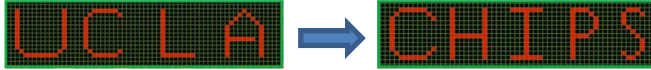

Fig. 5. Bitmaps of a fully functional CTT memory array integrated in 14nm FinFET technology: A "UCLA" pattern was written followed by an erase and re-write of a "CHIPS" pattern.

III. Program And Erase Optimization and Cycling

To demonstrate the importance of optimizing program and erase conditions for the memory application of CTT devices as discussed in section II, devices were cycled 20× using unoptimized P/E conditions (i.e. P/E conditions were not optimized to avoid over-programming and under-erasing) as well as optimized P/E conditions (V_{d-PRG} slightly lower than V_{d-INIT} was used and the number of 'PRG' pulses was limited to avoid over-programming. Longer erase times were 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(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(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. 5 shows bitmaps of a CTT memory array fabricated in 14nm FinFET production technology. Details about circuit design and sensing techniques implemented in this technology have been discussed elsewhere [12].

IV. Retention and Reliability

We would like to re-emphasize that charge trapping at (self-heating induced) high temperature is the key to the stability of the trapped charge [1]. As demonstrated and discussed in [1], devices programmed at higher V_d values show considerably higher charge retention as compared to devices programmed at lower V_d values (<10% charge loss after 10 years at 85°C for programming with $V_d = 1.3$V was demonstrated). Here, we quantify the impact of device self-heating during programming by measuring the activation energies (E_a) for charge de-trapping as a function of programming V_d and self-heating temperature. We monitored the reduction in ΔV_T of devices that were programmed using

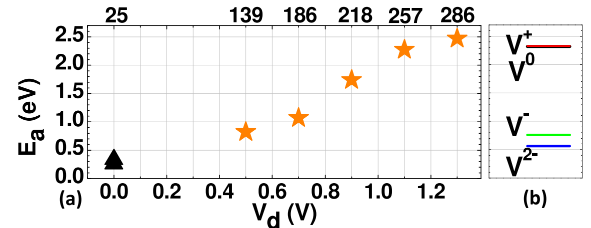

Fig. 6. (a) Measured activation energies (E_a) for charge de-trapping after programming at various V_d values (stars). Estimated channel temperatures (in °C) due to device self-heating during programming are indicated on the top scale. Measured E_a values (triangles) for de-trapping after trap filling during PBTI at room temperature are shown for comparison [14]. (b) Calculated thermal activation energies for de-trapping for various charge states of V_O in crystalline m-HfO_2 [15], revealing values ranging from 0.56eV–2.33eV.

various fixed V_d values and stored at various fixed elevated bake temperatures. A 'retention time' criteria of 15% ΔV_T degrade ($t_r^{15\%}$) was used and the E_a corresponding to each programming V_d was extracted from an Arrhenius plot of $t_r^{15\%}$, a method commonly used in literature [13], [14]. The results (Fig. 6(a)) clearly show that stability of the trapped charge is significantly enhanced by programing at high V_d values (or high self-heating temperatures), confirming the speculation in [1]. 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 de-trapping has been discussed in previous literatures [7], [11], [15], [16]. The variation in capture and emission times of HfO_2 traps has also been directly correlated to the spread in their activation energies [1], [17]. The calculated thermal activation energies for oxygen vacancy in its various charge states in crystalline m-HfO_2 is summarized in Fig. 6(b). In amorphous HfO_2, such energy levels are significantly spread in energy [17]. 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) [1]. Such traps are also likely to be more stable, resulting in a higher effective activation energy for de-trapping and enhanced stability (retention) of the CTT memory element.

V. Summary And Conclusions

In this letter, we have demonstrated a unique embedded non-volatile multiple-time programmable memory (MTPM) for SoC applications in advanced CMOS technology nodes, using HfO_2 based HKMG devices called CTTs, which has clear advantages over existing OTP memory technologies like the eFUSE [2] and gate breakdown anti-fuse [3]. This memory offers multiple-time programmability, low voltage and low power operation, high density, scalability, excellent retention, and can be integrated without any added process complexity or masks. Integration of functional memory arrays has been realized on 22nm planar and 14nm FinFET production technology platforms, demonstrating the robustness and commercial potential of this technology. Work to further model and optimize device and memory array design and operating conditions for this eNVM technology is ongoing.

We thank DARPA (award FA 8650-16-C-7648) and GlobalFoundries (especially Toshiaki Kirihata, Norman Robson, and Daniel Berger) for supporting this work. We also thank Xuefeng Gu of UCLA for fruitful conversations.

References

[1] 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 Lett.*, vol. 37, no. 1, pp. 88–91, Jan. 2016.

[2] 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.

[3] 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," in *Symp. VLSI-Technol. Dig.*, 2014, pp. 1–2.

[4] T. Hook, F. K. Allibert, Balakrishnan, B. Doris, D. N. Guo, Mavilla, E. Nowak, G. Tsutsui, R. Southwick, J. Strane, and X. Sun, "SOI FinFET versus bulk FinFET for 10nm and below," in *Proc. IEEE S3S Conf.*, Oct. 2014, pp. 1–3.

[5] 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," in *Proc. IEEE IEDM*, Dec. 2015, pp. 6–11.

[6] 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," in *Proc. IEDM*, 2012, pp. 52–55.

[7] 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," in *Proc. IEDM*, 2006, pp. 1–4.

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

[9] 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," in *Proc. IEEE IRPS*, Apr. 2009, pp. 486–492.

[10] 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," in *Proc. IRPS*, 2004, pp. 181–187.

[11] E. P. Gusev and C. P. D'Emic, "Charge detrapping in HfO_2 high-κ gate dielectric stacks," *Appl. Phys. Lett.*, vol. 83, no. 25, pp. 5223–5225, 2003.

[12] 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," in *Proc. Symp. VLSI-Circuits*, Jun. 2016, pp. 1–2.

[13] 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," *Microelectron. Eng.*, vol. 86, pp. 1796–1803, Sep. 2009.

[14] G. Bersuker, J. Sim, C. Park, C. Young, S. Nadkarni, R. Choi, and B. 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, Jan. 2007.

[15] 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," *Appl. Phys. Lett.*, vol. 89, p. 082908, May 2006.

[16] K. Xiong, J. Robertson, M. C. Gibson, and S. J. Clark, "Defect energy levels in HfO_2 high-dielectric-constant gate oxide," *Appl. Phys. Lett.*, vol. 87, p. 183505, Dec. 2005.

[17] T. Grasser, P.-J. Wagner, H. Reisinger, T. H. Aichinger, G. Pobegen, M. Nelhiebel, and B. Kaczer, "Analytic modeling of the bias temperature instability using capture/emission time maps," in *Proc. IEDM*, 2011, pp. 4–27.