Fully-CMOS Multi-Level Embedded Non-Volatile Memory Devices With Reliable Long-Term Retention for Efficient Storage of Neural Network Weights

Siming Ma, Marco Donato, Sae Kyu Lee, David Brooks, and Gu-Yeon Wei

***Abstract*— We present a fully CMOS-compatible multi-level non-volatile memory technology, without any special process cost. It is especially suitable for storing the weights of artificial neural networks on chip with low cost, high density, and high power-efficiency. We use hot carrier injection to program the single-transistor cells, and we conduct charge pumping experiments which identify interfacial traps, rather than bulk oxide traps, as the dominant factor in producing stable *I–V* shifts. We also derive a new physics-based experimentally verified logarithmic model to explain the rate of interfacial trap generation in large *I–V* shift regimes where the conventional power-law no longer applies. We fabricate two chips, one using TSMC's 16 nm FinFET and the other in 28 nm planar, and show the FinFET cells are more favorable for non-volatile memory due to their better channel control. We store multiple levels in each FinFET cell using a "program and check" strategy which sets memory cells' currents with standard deviations less than 2 μA across a shifting range of over 100 μA. We demonstrate 8 level FinFET cells with extrapolated 10-year charge loss within 10% at 125 °C.**

***Index Terms*— Embedded non-volatile memory, multi-level cells, artificial neural networks, MOSFET memory, hot carrier injection, interfacial traps.**

I. Introduction

ARTIFICIAL neural networks (ANN) have achieved amazing performance in various machine learning tasks [1]–[4]. However, state-of-the-art ANNs have huge numbers of parameters (mostly weights), easily exceeding $10^7 \sim 10^8$ [1], [5], [6]. When deploying these ANNs on hardware for inference, it is impractical to fit all the weights in on-chip SRAMs, so they have to access slow and power consuming off-chip DRAMs.

After an ANN is trained, the weights only need to be written once and are fixed during inference. Moreover, ANNs are known for their noise resilience, which provides the opportunity to trade off the noise of the weights for their storage efficiency. Embedded non-volatile memories (eNVM) are therefore proposed to replace power and area consuming SRAMs for ANN weight storage, since eNVMs can have much higher density and lower leakage, while their long writing time is only a one-time cost. However, these eNVMs, which include embedded flash [7], phase change and resistive RAMs [8], require extra cost for their special process steps. On the other hand, recent research shows the potential of using purely-CMOS charge trap transistors for compact eNVMs [9]–[14]. In this letter, we present a fully CMOS-compatible multi-level eNVM technology, to enable high-density and low-cost weight storage for on-chip ANN inference.

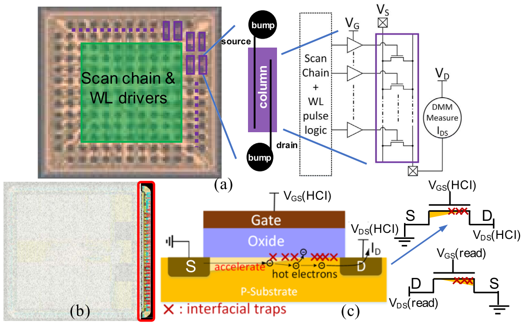

Fig. 1. Test chips in TSMC HKMG (a) 16nm FinFET, and (b) 28nm planar (chip area shared with another project, eNVM cells are highlighted on the right). In both chips, eNVM cells are organized as NOR-type columns whose source and drain terminals can be connected to an external digital multimeter (DMM) for current measurements. The device cross-section in (c) illustrates the physics of HCI during programming, which corresponds to the NMOS circuit symbol on the upper right. The source and drain are flipped during reading (as shown in the bottom NMOS circuit symbol), in order for the traps to maximally influence the inverted channel. Although here we draw a planar device for clear illustration, this physics of HCI also applies to FinFETs.

II. Test Chips and Hot Carrier Injection

We fabricate two test chips using TSMC's HKMG 16nm FinFET (Fig. 1a) and 28nm planar (Fig. 1b) processes [15], [16]. Both chips contain core NMOS transistors organized in NOR-type memory columns, whose source and drain terminals are externally driven, and the current of selected cell(s) is measured by a digital multimeter.

We use hot carrier injection (HCI) to program our eNVM cells. Although HCI is traditionally viewed as an aging effect [17], we intentionally accelerate HCI to induce drastic changes in transistors' I-V characteristics for long-term data storage. Under higher than nominal V_{DS} and V_{GS} biasing voltages, the high horizontal electric field accelerates channel electrons such that some electrons can no longer maintain thermal equilibrium with the silicon lattice and become

Manuscript received July 2, 2019; accepted July 17, 2019. Date of publication July 22, 2019; date of current version August 23, 2019. The review of this letter was arranged by Editor B. S. Doyle. *(Corresponding author: Siming Ma.)*

S. Ma, M. Donato, D. Brooks, and G.-Y. Wei are with Harvard University, Cambridge, MA 02138 USA (e-mail: simingma@g.harvard.edu).

S. K. Lee was with Harvard University, Cambridge, MA 02138 USA. He is now with IBM Research, Yorktown Heights, NY 10598 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.2019.2930212

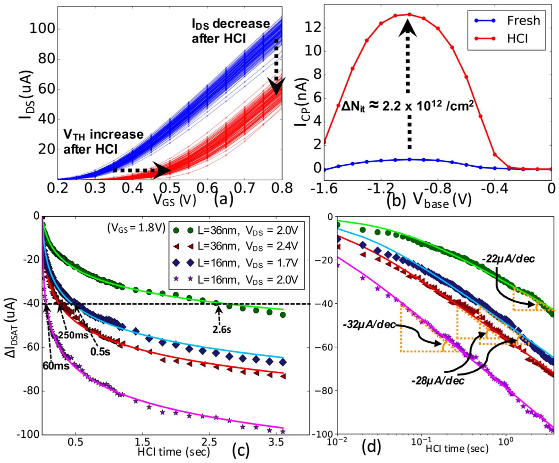

Fig. 2. (a) shows measured I-V curves of 128 NMOS cells (16nm FinFET, $L = 36nm$, 2 fins) before and after HCI, and (b) is the corresponding total charge pumping current (I_{cp}) of all these 128 cells with 5MHz pumping frequency [19], measured before and after HCI. (c) and (d) show reduction in I_{DSAT} (at $V_{DS} = V_{GS} = 0.8V$) vs HCI time (in linear-scale and log-scale, respectively), for different L and V_{DS} pairs (for each pair, we measure the average current across 16 devices in between HCI pulses). The lines correspond to curve fits of the logarithmic equation (2) to the measured data.

"hot-electrons" [18]. These "hot-electrons" can either break the passivation Si-H bonds at the oxide-silicon interface, creating interfacial traps, or surmount the oxide energy barrier and generate bulk oxide traps (Fig. 1c). These effects will cause the transistors' I_{DS}-V_{GS} curves to shift down, measuring a smaller read current under the same read voltage, thereby storing new data in a non-volatile fashion. Fig. 2a shows the shifts of I_{DS}-V_{GS} curves of 128 NMOS transistors (16nm FinFET process, $L = 36nm$, 2 fins) due to HCI with $V_{GS} = 1.8V$ and $V_{DS} = 2.0V$ for $2.4\,seconds$. We can distinguish two separate levels, even with the random variations within levels.

A. Interfacial Traps vs Bulk Oxide Traps

For the purpose of on-chip ANN weight storage, HCI-induced interfacial traps are superior to bulk oxide traps due to their larger I-V shifting range and more reliable retention. We conduct charge pumping experiments [20], which identify interfacial traps as the dominant factor in producing significant and reliable changes in transistors' I-V characteristics. We use the "constant magnitude pulse" technique [19] with 5MHz pumping frequency on all the 128 NMOS devices in Fig. 2a, and measure their aggregated charge pumping current (I_{cp}) before and after HCI (Fig. 2b). The maximum I_{cp} increases from $0.784nA$ to $13.157nA$, corresponding to an increase of interfacial trap density $Q_{it} \approx 2.2 \times 10^{12}/cm^2$, or a threshold voltage shift $\Delta V_{TH} = qQ_{it}/C_{ox} \approx 155mV$ (C_{ox} is the oxide capacitance per cm^2), which accounts for almost the entirety of measured ΔV_{TH}. This proves that interfacial traps, rather than bulk oxide traps, are what cause the stable I-V shifts.

We also discover the Fowler-Nordheim (FN) tunneling condition (applying a very high gate voltage $\sim 2.4V$ while grounding the source, drain, and body) can generate bulk oxide traps without creating interfacial traps. However, the maximum ΔV_{TH} due to these bulk oxide traps is less than $50mV$, insufficient to separate 2 levels. Due to the thinness of oxide in advanced CMOS processes, these bulk oxide traps fully de-trap after being left unbiased for a few minutes, or tunnel back to the body even faster under a negative gate bias.

Compared with bulk oxide traps, interfacial traps are more stable [21], and we demonstrate reliable shift from $I_{DSAT} >$ $120\mu A$ down to less than $5\mu A$ ($\Delta V_{TH} > 500mV$), without inducing oxide breakdown. Although interfacial traps cannot be erased using electrical field, prior work shows they can fully anneal within $100\,seconds$ at 380°C or $2\,hours$ at 200°C, for hydrogen to diffuse back and re-passivate the silicon dangling bonds [22]–[24]. Due to lack of local heaters in our current chips to generate high enough temperatures, these eNVM cells are currently suitable for well-trained ANNs that only need to be written once and are read-only during inference.

B. Programming Time

The programming time is very sensitive to HCI conditions, and higher voltages or shorter L can dramatically speed up writing. Previous HCI studies mostly focus on aging effects, with relatively small and gradual I-V shifts that can be reasonably modeled by the traditional power law [25]. However, we want to accelerate HCI to quickly achieve a large shift, which is well beyond the range the power law can model. Consistent with the "self-limiting effect" in severe HCI scenarios that others also observed [26], our experiments show a logarithmic progression of ΔI_{DSAT} vs HCI time, for which we derive a new physical model as follows. The generation rate of interfacial trap density Q_{it} follows the Arrhenius equation, with an activation energy E_a related to the Si-H bonding energy. Initially, the average hot electron energy E_{hci} is determined by the programming voltages and transistor geometry. As Q_{it} increases, the interfacial trapped electrons create an increasing Coulomb repelling force against hot electrons and reduce their average energy into $E_{hci} - \alpha Q_{it}$, where α is a constant of proportionality. The reaction rate is also proportional to the density of electron supply, I_{DS}/W (W is the effective channel width in the case of FinFET), and the density of available Si-H bonds, which is very high ($\sim 10^{20}cm^{-3}$) [17], therefore considered as constant and absorbed into the pre-exponential factor k:

$$\begin{aligned}\frac{dQ_{it}}{dt} &= k\frac{I_{DS}}{W}\exp(-\frac{E_a}{E_{hci} - \alpha Q_{it}}) \\ &\approx k\frac{I_{DS}}{W}\exp[-\frac{E_a}{E_{hci}}(1 + \frac{\alpha Q_{it}}{E_{hci}})] \qquad (1)\end{aligned}$$

The approximation uses Taylor expansion and assumes $\alpha Q_{it} \ll E_{hci}$. Separating Q_{it} and t into two sides of the equation and integrating once gives: $Q_{it}(t) = A\ln(1 + \frac{R_0}{A}t)$, where $A = \frac{E_{hci}^2}{\alpha E_a}$, and $R_0 = k\frac{I_{DS}}{W}\exp(-\frac{E_a}{E_{hci}})$ is the initial rate $\frac{dQ_{it}}{dt}|_{t=0}$. $Q_{it}(t)$ increases V_{TH} by $\Delta V_{TH}(t) = qQ_{it}(t)/C_{ox}$, and since velocity saturation results in a linear I-V dependence: $I_{DSAT} = WC_{ox}(V_{GS} - V_{TH})\upsilon_{sat}$ (υ_{sat} is the saturation velocity) [27], ΔI_{DSAT} is linearly dependent on Q_{it}:

$$\begin{aligned}&\Delta I_{DSAT}(t) \\ &= -WC_{ox}\upsilon_{sat}\Delta V_{TH}(t) = -W\upsilon_{sat}qA\ln(1 + \frac{R_0}{A}t) \qquad (2) \\ &\approx \begin{cases} -B\cdot t, & \text{when } \frac{R_0}{A}t \ll 1 \qquad (3a) \\ -W\upsilon_{sat}qA\ln(\frac{R_0}{A}) - C\log(t), & \text{when } \frac{R_0}{A}t \gg 1 \qquad (3b)\end{cases}\end{aligned}$$

where $B = W\upsilon_{sat}qR_0$ and $C = \frac{W\upsilon_{sat}qA}{\log e}$. Shown in Fig. 2c and 2d, the logarithmic model in (2) matches data well, and V_{DS} and L heavily impact ΔI_{DSAT} trajectories through their influence on E_{hci}. At the beginning of HCI ($\frac{R_0}{A}t \ll 1$ as in (3a)), the Coulomb repulsion of Q_{it} is negligible, so I_{DSAT} decreases linearly w.r.t. t with a rate B. We find B to vary across a wide range $\sim 10^2 \sim 10^4$ ($\mu A/sec$) in different HCI conditions, due to its strong dependence on E_{hci} through the exponential term in R_0. After Q_{it} is large enough to build up significant Coulomb field, I_{DSAT} reduces linearly w.r.t. $\log(t)$

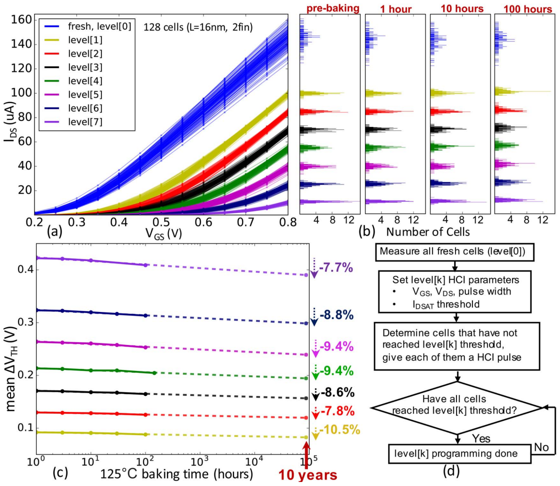

Fig. 3. (a) shows the measured I-V curves ($V_{DS} = 0.8V$) of 128 MLCs with 8 levels (3 bits) of storage. (b) shows the corresponding distributions of all the 8 levels of I_{DSAT} (at $V_{GS} = V_{DS} = 0.8V$) when baked at 125°C for up to 100 *hours*, confirming stable long-term retention, with I_{DSAT} standard deviations of all the programmed levels maintained within $2\mu A$. (c) plots the mean ΔV_{TH} vs 125°C baking time, showing that the extrapolated 10-year charge losses of all the programmed levels are within 10%. (d) is a flow chart of the iterative "program and check" procedure to program from the fresh cells to a certain level[*k*].

as in (3b), decreasing by $C(\mu A)$ per decade increase of t. E_{hci} impacts C through the quadratic term in A, which is a weaker relationship compare to R_0, and we find C to be $\sim 20 \sim 40\mu A/dec$ across various HCI conditions. Therefore, a larger E_{hci} gives a larger $\frac{R_0}{A}$, so that it takes a shorter t for ΔI_{DSAT} to reach the log-shift regime in (3b). As labeled in Fig. 2c, by using a shorter L and larger V_{DS} to increase E_{hci}, we can reduce the HCI time to achieve $|\Delta I_{DSAT}| = 40\mu A$ from $2.6s$ to $60ms$. Shown in Fig. 2d with t in log-scale, after $|I_{DSAT}|$ is large enough (about $> 40\mu A$), all the curves transition to straight lines following (3b). Our multi-level programming strategy in III-A also relies on this model.

C. FinFET Cells vs Planar Cells

HCI programming works for both the 16nm FinFET and 28nm planar cells. However, we find the FinFET transistors preferable due to their superior channel control. The source and drain are shared among the cells in a memory column, so when writing a cell by high V_{DS} and V_{GS}, other unselected cells ($V_{GS} = 0$) on the same column still see the high V_{DS}. For planar cells with $L = 30nm$ (L_{min} in this process), we observe severe punch-through current [28] in unselected cells when $V_{DS} > 2V$, which creates hot electrons and causes unintentional programming (write disturb). To lessen the punch-through effect, we have to use smaller V_{DS} and/or longer L, which increases the writing time and/or hurts the memory density. In comparison, the FinFET cells have better channel control, and we do not observe punch-through induced write disturb using $L = 16nm$ with V_{DS} up to $2.3V$. The FinFET cells are therefore preferred, as they operate reliably without sacrificing storage density or writing speed.

III. Multi-Level Cells (MLC)

A. MLC Programming Strategies

Interfacial traps enable wide I-V shifting ranges for storing multiple levels, but the main challenge is to tighten the level distributions seen in Fig. 2a, to maintain sufficient inter-level margins which affect the read error rate. We design an iterative "program and check" procedure shown in Fig. 3d, wherein we progressively stress each transistor with discrete HCI pulses until it reaches a desired I_{DSAT} threshold of a programming level. This method uses variable numbers of pulses for different cells to compensate for process variations and the randomness of HCI. Care should be taken on the short-term relaxation effect due to small amounts of bulk oxide traps that can change I_{DSAT} by up to a few μAs. Consistent with the discussions in II-A, we find that larger V_{GS} causes more bulk oxide trapping due to FN-tunneling near the source. These bulk oxide trapped electrons escape in a few minutes, so the timing sequence should allow this short-term relaxation to finish before checking I_{DSAT} to decide whether a cell has reached the target I_{DSAT} threshold or not.

The strategy of choosing HCI pulse widths leverages the logarithmic model derived in II-B. For sufficient inter-level margins, we define all of the programming levels to be in the log-shift regime in (3b) with an equal I_{DSAT} distance (denoted as ΔI) between adjacent levels. The expected total HCI times to reach any 2 adjacent levels (denoted as T_l and T_{l+1} for levels l and $l+1$) have a constant ratio $\gamma = \frac{T_{l+1}}{T_l} = 10^{\frac{\Delta I}{C}}$ (since $C\log(T_{l+1}) - C\log(T_l) \approx \Delta I$ from (3b)). Taking the derivative of (2), the rate of I_{DSAT} shift is $\frac{d\Delta I_{DSAT}(t)}{dt} = -\frac{W v_{sat} q R_0}{1+\frac{R_0}{A}t} \approx \frac{W v_{sat} q A}{t}$ (when $\frac{R_0}{A}t \gg 1$), meaning that the expected I_{DSAT} shift during a short pulse width δt_l at the vicinity of level[l]'s threshold is $\delta I_l \approx W v_{sat} q A \frac{\delta t_l}{T_l}$. Therefore, we can scale up δt_l with the same ratio γ while still maintaining similar I_{DSAT} tightness in all the programmed levels, since the spread of I_{DSAT} for level[l] is mostly due to the random overshoot of δI_l during the δt_l pulses.

As an example of applying these MLC programming strategies, we demonstrate 8-level cells, defining the 7 programming level I_{DSAT} thresholds to be from $100\mu A$ down to $10\mu A$ with an equal distance $\Delta I = 15\mu A$. Using the fastest programming condition in Fig. 2c and 2d ($L = 16nm$ FinFETs using $V_{GS} = 1.8V$ and $V_{DS} = 2.0V$, with $C \approx 32\mu A/dec$), we determine $\gamma = \frac{T_{l+1}}{T_l} = \frac{\delta t_{l+1}}{\delta t_l} = 10^{\frac{15}{32}} \approx 3$. Choosing $\delta t_1 = 1ms$ that achieves sufficiently small overshoot for level[1], we can determine the pulse widths for the rest of programming levels $\delta t_2 \sim \delta t_7$ to be $3ms, 9ms, \ldots, 729ms$. Fig. 3a shows the 8-level experiment results for a 128-cell column ($L = 16nm$, 2 fins), which achieves 8 distinctly separated levels by tightening the level distributions and taking advantage of the entire current range. The distributions of I_{DSAT} (measured at $V_{GS} = V_{DS} = 0.8V$) of all the programmed levels indeed have similar tightness, with an average standard deviation of $1.3\mu A$.

B. MLC Retention

We verify the reliable retention of MLCs by baking the test chip in Fig. 3a at 125°C for 100 *hours* for all the levels. Shown in Fig. 3b, all the level distributions retain tightness, with I_{DSAT} standard deviations of all the programmed levels maintained within $2\mu A$. Extrapolating their ΔV_{TH} retention at 125°C to 10 *years* (Fig. 3c), the charge losses of all the programmed levels are within about 10%, which is sufficiently stable for a variety of ANN applications.

Acknowledgment

We thank Professor Evelyn Hu for insightful discussions.

References

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in *Advances in Neural Information Processing Systems*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. New York, NY, USA: Curran Associates, Inc., 2012, pp. 1097–1105. [Online]. Available: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

[2] I. Sutskever, O. Vinyals, and Q. V. Le, “Sequence to sequence learning with neural networks,” in *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. New York, NY, USA: Curran Associates, Inc., 2014, pp. 3104–3112. [Online]. Available: http://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.pdf

[3] A. Graves, A.-R. Mohamed, and G. Hinton, “Speech recognition with deep recurrent neural networks,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 6645–6649. doi: 10.1109/ICASSP.2013.6638947.

[4] A. Y. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, and A. Y. Ng, “Deep speech: Scaling up end-to-end speech recognition,” *CoRR*, vol. abs/1412.5567, pp. 1–12, Dec. 2014. [Online]. Available: https://arxiv.org/abs/1412.5567

[5] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *CoRR*, vol. abs/1409.1556, pp. 1–14, Sep. 2014. [Online]. Available: https://arxiv.org/abs/1409.1556

[6] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[7] L. Fick, D. Blaauw, D. Sylvester, S. Skrzyniarz, M. Parikh, and D. Fick, “Analog in-memory subthreshold deep neural network accelerator,” in *Proc. IEEE Custom Integr. Circuits Conf. (CICC)*, Apr./May 2017, pp. 1–4. doi: 10.1109/CICC.2017.7993629.

[8] Y. Fujisaki, “Review of emerging new solid-state non-volatile memories,” *Jpn. J. Appl. Phys.*, vol. 52, no. 4R, 2013, Art. no. 040001. [Online]. Available: http://stacks.iop.org/1347-4065/52/i=4R/a=040001

[9] 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,” in *Proc. IEEE Int. Rel. Phys. Symp.*, Apr. 2015, pp. MY.2.1–MY.2.4. doi: 10.1109/IRPS.2015.7112816.

[10] 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, “80 Kb 10 ns read cycle logic embedded high-*k* charge trap multi-time-programmable memory scalable to 14 nm FIN with no added process complexity,” in *Proc. IEEE Symp. VLSI Circuits (VLSI-Circuits)*, Jun. 2016, pp. 1–2. doi: 10.1109/VLSIC.2016.7573462.

[11] 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-κ-metal-gate nFETs,” *IEEE Electron Device Lett.*, vol. 37, no. 1, pp. 88–91, Jan. 2016. doi: 10.1109/LED.2015.2504952.

[12] 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 Lett.*, vol. 38, no. 1, pp. 44–47, Jan. 2017. doi: 10.1109/LED.2016.2633490.

[13] B. Jayaraman, D. Leu, J. Viraraghavan, A. Cestero, M. Yin, J. Golz, R. R. Tummuru, R. Raghavan, D. Moy, T. Kempanna, F. Khan, T. Kirihata, and S. S. Iyer, “80-kb logic embedded high-*k* charge trap transistor-based multi-time-programmable memory with no added process complexity,” *IEEE J. Solid-State Circuits*, vol. 53, no. 3, pp. 949–960, Mar. 2018. doi: 10.1109/JSSC.2017.2784760.

[14] E. Hunt-Schroeder, D. Anand, J. Fifield, M. Jacunski, M. Roberge, D. Pontius, K. Batson, and T. Kirihata, “14 nm FinFET 1.5 mb embedded high-*k* charge trap transistor one time programmable memory using dynamic adaptive programming,” in *Proc. IEEE Symp. VLSI Circuits*, Jun. 2018, pp. 87–88. doi: 10.1109/VLSIC.2018.8502415.

[15] S. Y. Wu, C. Y. Lin, M. C. Chiang, J. J. Liaw, J. Y. Cheng, S. H. Yang, S. Z. Chang, M. Liang, T. Miyashita, C. H. Tsai, C. H. Chang, V. S. Chang, Y. K. Wu, J. H. Chen, H. F. Chen, S. Y. Chang, K. H. Pan, R. F. Tsui, C. H. Yao, K. C. Ting, T. Yamamoto, H. T. Huang, T. L. Lee, C. H. Lee, W. Chang, H. M. Lee, C. C. Chen, T. Chang, R. Chen, Y. H. Chiu, M. H. Tsai, S. M. Jang, K. S. Chen, and Y. Ku, “An enhanced 16 nm CMOS technology featuring 2^{nd} generation FinFET transistors and advanced Cu/low-k interconnect for low power and high performance applications,” in *IEDM Tech. Dig.*, Dec. 2014, pp. 3.1.1–3.1.4. doi: 10.1109/IEDM.2014.7046970.

[16] S. H. Yang, J. Y. Sheu, M. K. Ieong, M. H. Chiang, T. Yamamoto, J. J. Liaw, S. S. Chang, Y. M. Lin, T. L. Hsu, J. R. Hwang, J. K. Ting, C. H. Wu, K. C. Ting, F. C. Yang, C. M. Li, I. L. Wu, Y. M. Chen, S. J. Chent, K. S. Chen, J. Y. Cheng, M. H. Tsai, W. Chang, R. Chen, C. C. Chen, T. L. Lee, C. K. Lin, S. C. Yang, Y. M. Sheu, J. T. Tzeng, L. C. Lu, S. M. Jang, C. H. Diaz, and Y. Mii, “28 nm metal-gate high-*k* CMOS SoC technology for high-performance mobile applications,” in *Proc. IEEE Custom Integr. Circuits Conf. (CICC)*, Sep. 2011, pp. 1–5. doi: 10.1109/CICC.2011.6055355.

[17] C. Hu, S. C. Tam, F.-C. Hsu, P.-K. Ko, T.-Y. Chan, and K. W. Terrill, “Hot-electron-induced MOSFET degradation—Model, monitor, and improvement,” *IEEE Trans. Electron Devices*, vol. 32, no. 2, pp. 375–385, Feb. 1985. doi: 10.1109/T-ED.1985.21952.

[18] C. C. Hu, “Lucky-electron model of channel hot electron emission,” in *IEDM Tech. Dig.*, Dec. 1979, pp. 22–25. doi: 10.1109/IEDM.1979.189529.

[19] A. B. M. Elliot, “The use of charge pumping currents to measure surface state densities in MOS transistors,” *Solid State Electron*, vol. 19, no. 3, pp. 241–247, Mar. 1976. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0038110176901696. doi: 10.1016/0038-1101(76)90169-6.

[20] P. Heremans, J. Witters, G. Groeseneken, and H. E. Maes, “Analysis of the charge pumping technique and its application for the evaluation of MOSFET degradation,” *IEEE Trans. Electron Devices*, vol. 36, no. 7, pp. 1318–1335, Jul. 1989. doi: 10.1109/16.30938.

[21] S. Mahapatra, D. Saha, D. Varghese, and P. B. Kumar, “On the generation and recovery of interface traps in MOSFETs subjected to NBTI, FN, and HCI stress,” *IEEE Trans. Electron Devices*, vol. 53, no. 7, pp. 1583–1592, Jul. 2006. doi: 10.1109/TED.2006.876041.

[22] G. Pobegen, S. Tyaginov, M. Nelhiebel, and T. Grasser, “Observation of normally distributed energies for interface trap recovery after hot-carrier degradation,” *IEEE Electron Device Lett.*, vol. 34, no. 8, pp. 939–941, Aug. 2013. doi: 10.1109/LED.2013.2262521.

[23] J.-W. Han, M. Kebaili, and M. Meyyappan, “System on microheater for on-chip annealing of defects generated by hot-carrier injection, bias temperature instability, and ionizing radiation,” *IEEE Electron Device Lett.*, vol. 37, no. 12, pp. 1543–1546, Dec. 2016. doi: 10.1109/LED.2016.2616133.

[24] J.-W. Han, R. Peterson, D.-I. Moon, D. G. Senesky, and M. Meyyappan, “Monolithically integrated microheater for on-chip annealing of oxide defects,” *IEEE Electron Device Lett.*, vol. 38, no. 7, pp. 831–834, Jul. 2017. doi: 10.1109/LED.2017.2700326.

[25] E. Takeda and N. Suzuki, “An empirical model for device degradation due to hot-carrier injection,” *IEEE Electron Device Lett.*, vol. EDL-4, no. 4, pp. 111–113, Apr. 1983. doi: 10.1109/EDL.1983.25667.

[26] C. Liang, H. Gaw, and P. Cheng, “An analytical model for self-limiting behavior of hot-carrier degradation in 0.25 μm n-MOSFET’s,” *IEEE Electron Device Lett.*, vol. 13, no. 11, pp. 569–571, Nov. 1992. doi: 10.1109/55.192843.

[27] B. G. Streetman and S. Banerjee, *Solid State Electronic Devices*, vol. 4. Englewood Cliffs, NJ, USA: Prentice-Hall 1995.

[28] W.-S. Feng, T. Y. Chan, and C. Hu, “MOSFET drain breakdown voltage,” *IEEE Electron Device Lett.*, vol. EDL-7, no. 7, pp. 449–450, Jul. 1986. doi: 10.1109/EDL.1986.26432.