Received 10 May 2020; accepted 26 June 2020. Date of publication 30 June 2020; date of current version 14 July 2020.
The review of this article was arranged by Editor M. Liu.

*Digital Object Identifier 10.1109/JEDS.2020.3005904*

# Self-Clamping Programming in Narrow-Bridge Floating Gate Cells for Multi-Level Logic Non-Volatile Memory Applications

**WEI-CHENG ZHUANG, CHING-TING CHIEN (Member, IEEE), CHRONG JUNG LIN (Associate Member, IEEE), AND YA-CHIN KING (Senior Member, IEEE)**

Institute of Electronics Engineering, National Tsing Hua University, Hsinchu 300, Taiwan

CORRESPONDING AUTHOR: W.-C. ZHUANG (e-mail: lmh830231@gmail.com)

This work was supported by Vanguard International Semiconductor (VIS).

**ABSTRACT** A new self-converging programming characteristic in a single-poly floating-gate memory cell with full-compatibility to a CMOS logic technology is observed and studied. A uniquely design cell with a narrow-bridging line between two coupling capacitors promotes a localized charging effect at the electron tunneling site, leading to clamping of threshold voltage states. Through this mechanism, the new multi time programmable (MTP) cells exhibit tight threshold voltage distributions for multi-level cells (MLC) operations. Improved cycling reliability and one-shot multi-level programming has been fully demonstrated in this work.

**INDEX TERMS** MTP, MLC, flash memory, logic NVM.

## I. INTRODUCTION

As the market demand for embedded memory grows, there are many solutions that serve different users meeting their various needs on versatile CMOS platforms. As the technology scales, changes in material/film thickness/structure the making of CMOS circuits putting constraints on the design on floating gate (FG) logic non-volatile memory cells, causing limitations on their cell size, performance and reliability [1]. MLC [2]–[4] are one of the common and effective schemes to raise storage volume without changing the basic memory array hardware. This method has been extended to triple-level cell (TLC) and quadruple-level cell (QLC) technologies used in many commercial non-volatile memory products [5]. Although this method can greatly enhance memory cell density, storage data integrity is greatly compromised, as a consequence [6]–[8]. In addition to MLC technology, where the discrete threshold voltage levels are used for multi-level representation, a charge-level control method has also been reported, as studied in rank modulation scheme [9]. As reported in [10], with the optimized approach applied, this scheme can not only increase the data storage capacity, but also reduce the issue of the overlapping of threshold levels. Unfortunately, to distinguish the multi-bit stored data, complex read operations are needed in such schemes, which in turn decreases the operation efficiency. For MLC applications, the optimized programming operation are critical to the final implementation of such schemes, hence are studied, extensively [10]–[12]. For multi-level bit per cell memories, built-in program-verify loops [13] are essential to prevent overlapping of the states, tighten threshold voltage distributions for cells across a gigabit memory array. To ensure MLC operation, writing of data generally requires going into a few program-verify iterations [14], which in-turn calls for extra circuits, e.g., error correcting code (ECC) [15] and redundancy [16]–[17], increase the overhead on peripheral circuits. Indispensable complex programming cycles can slow down programming speed, increases operation power and might induce long-term reliability issues on data integrity [18]. The read speed of single-level cells (SLC) and MLC devices are almost identical while MLC is almost 3∼4 times slower in terms of write performance as compared to SLC [19]. To enhance cell performance whiling maintaining full compatibility to standard CMOS processes, hybrid storage structure, which combine a few SLC cells along with the MLC devices are proposed [20]. In this study, with narrowing of the extended floating gate, a distinct “saturation voltage level” was found in unique cells with. When removing stored electrons from the FG, a localized charging effect was found, causing a transient saturation of the threshold voltage. This unique feature is applied to the programming of multiple-level storage on the narrow-bridge single-poly MTP cells. The newly discovered unique programming characteristics on these cells, fully-compatible to standard CMOS process, is studied for precise $V_{th}$ level control in multi-level cell operations. This cell with proper operation scheme provides an easier mechanism to obtained stable multi-level threshold voltage states. Comparing to the previously reported single-poly MTP cells [21]–[22], this new cell achieves multi-level cell (MLC) for higher density data storage. In addition, lower power consumption and enhances programming efficiency can be obtained with a single-pulse operation as compared to the conventional MLC methods [23]–[25].

## II. CELL STRUCTURE AND OPERATION PRINCIPLE

The narrow-bridging MTP devices studied here are designed and implemented using an 0.18µm CMOS process. The MTP memory are generally referred to the type of embedded non-volatile memories which can be reprogrammed and updated for a few thousand times [26]. The CMOS technology used to implement the memory cells has a critical dimension of 0.18µm, which directly correlate to the minimum gate length of a transistor. This cell however can be extended to other advanced CMOS technology nodes with provide interface devices operating at a supply voltage level of 3.3V. Fig. 1(a) is the 3D illustration of the proposed single-poly floating gate cell structure, with its circuit symbol in Fig. 1(b). A polysilicon FG is laid on top of two isolated n-well regions as the program gate (PG) and the erase gate (EG), respectively. A read transistor channel is placed in between the two capacitors. The narrow FG bridging the two capacitors is 360nm in width and 4.92µm in length, which is found to amplify the localized charging effect when electrons are pulling out from FG. Based on previous work and the following experiment results, it apparently shows that the localized charging effect need two conditions. First, the oxide thickness couldn’t be more than 70Å. (The dielectric thickness of [21] is 110 Å.) Second, according to the measurement result of the identical structure used 28nm process. Due to the technology change of gate dielectric from poly gate to metal gate, the resistance of FG significantly reduces. Hence, the resistance between the strong control gate (sCG) and weak control gate (wCG) should be higher than a specific level to ensure the delay occurred. Fig. 1 (c) shows the cross-sectional view of an unit cell with two well coupling nodes. The top-view and its equivalent circuit model are shown in Fig. 2. A resistor links the sides coupled by the wCG and the sCG, respectively. This implies that significant RC delay can affect the transient response before charge reach a steady state. When a high voltage is applied on wCG, electrons are pull out from the FG by FN tunneling, see Fig. 2 (b). Electron tunneling from FG-wCG follows the FN tunneling mechanism, which required a voltage difference between the two terminals exceeding 3.2eV [27]. With a tunneling oxide thickness of 7nm for our samples, an electric field of 13MV/cm. sufficient to induced a FN tunneling at a level 20nA at the FG tip. However, the narrow-bridge between node X and Y builds a barrier for the FG to reach charge equilibrium quickly. In a narrow FG, time for reaching charge equilibrium between the two capacitors increases, leading to localized charge (electrons) remain at node Y. This cause positive charge accumulating at node X, as resistance linking the two capacitances increases. As a result, the potential difference between FG and wCG decreases, significantly reduces the FN tunneling probability. During the removal of electrons in the FG from under the wCG, measured data in Fig. 3 indicate the threshold voltage holds at several saturation levels rather than continuously dropping of threshold states. The level of the saturation state is defined as “saturated threshold level”, while Δt specifies the minimum pulse width that triggered the localized charging effect. Data in Fig. 3 also suggests that voltage on wCG affects to both Vth level and time-to-reach each saturation stages.

The saturation levels change with increasing VwCG are summarized in Fig. 4(a). As expected, raising VwCG required the potential at node X to be lowered to stop FN tunneling from happening. Hence, results in a lower saturation

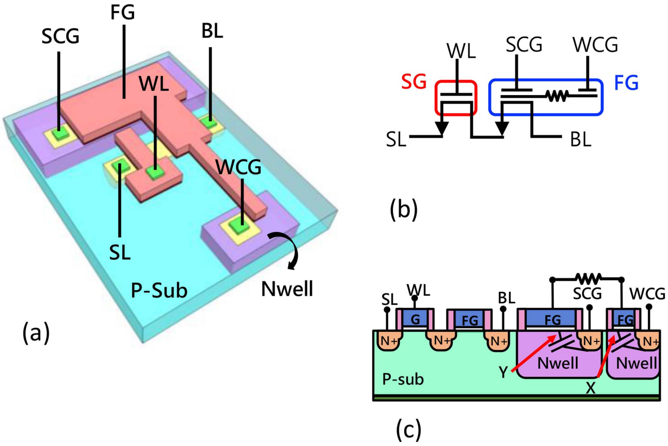


**FIGURE 1.** **(a) Narrow-bridge Floating gate (FG) cell in 3D illustration, (b) its equivalent circuit symbol for the cell and (c) the cross-section view of the cell.**

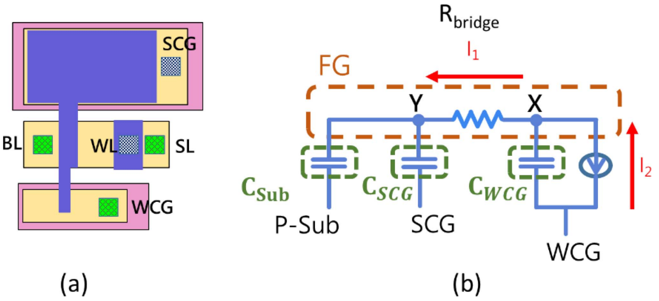


**FIGURE 2.** **(a) Layout of the proposed cell and (b) the equivalent circuit model for simulating the localized charge effect.**

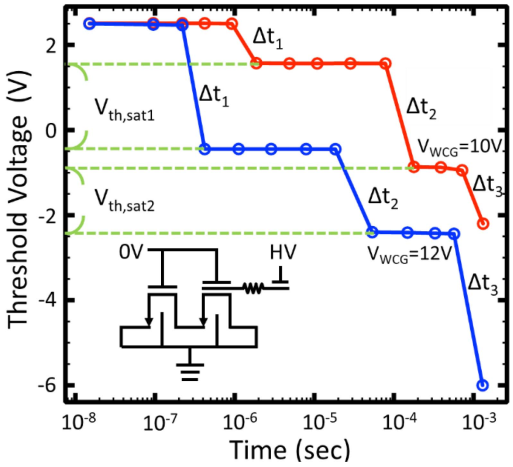
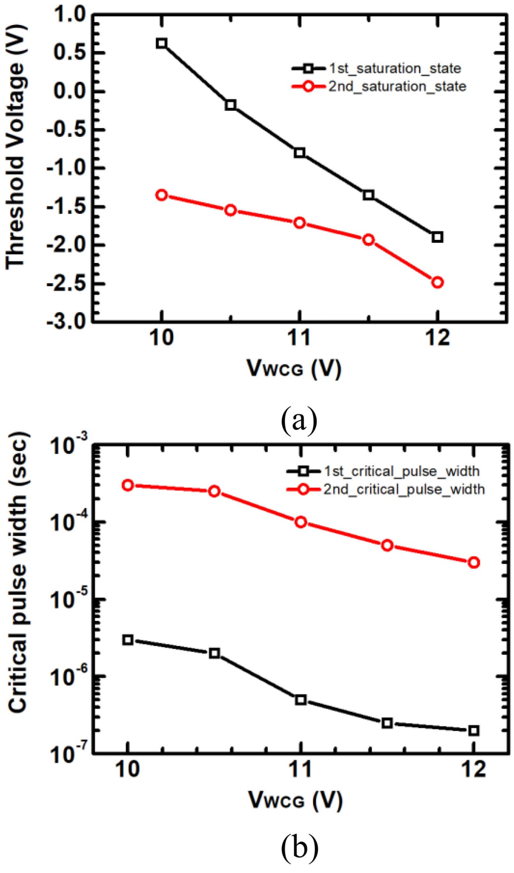


**FIGURE 3.** The measurement result of staurated $V_{th}$ levels when electrons are pulled out of the floating gate.



(a)

(b)

**FIGURE 4.** The saturation threshold levels at different stages and the corresponding critical pulse width with increasing $V_{wCG}$ levels.

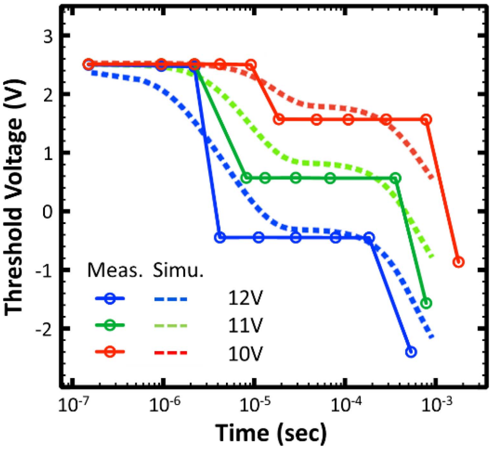


**FIGURE 5.** Comparison of the measurement results with the simulated RC effect on the narrow-bridge floating gate structure with $R_{bridge}$ = 1kΩ, $\frac{C_{WCG}}{C_{SCG}+C_{Sub}}$=20, where the solid-line are measured data, dash-lines are simulated data.

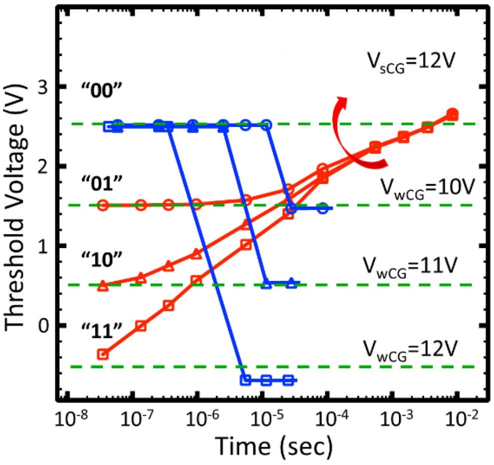


**FIGURE 6.** Time to program/erase characteristics of the MLC cells.

level. Fig. 4(b) compared the time required to have the significant charge build up at node X, Δt vs. VwCG, which also suggest that high voltage increase the speed of the local charging effect. Experiment result suggested that strong positive correlation between the saturated threshold voltage and VwCG. This phenomenon can be further explained by incorporating the RC equivalent model within the narrow-bridge FG structure during electron removal. Through simulation based on this model, one can accurately predict the temporary clamping characteristics between threshold voltage and applied pulse width in Fig. 5. As suggested, the number of electrons being pull out before saturation occurs is decided by the average FN current level and Δt. While giving a wider pulse shift the saturation levels into different stages, the saturated Vth levels shift further down with increasing VwCG, which suggests that multi-level Vth states can be achieved through VwCG control rather than pulse width (or accumulated number of pulse) in conventional multi-level cells. The time-to-program from initial states to other three different level for 2-bit per cell storage and the time-to-erase characteristics from three different states are compared in Fig. 6. The new voltage level control scheme for reaching precise multi-level Vth is proposed and demonstrated, here. Four states can be easily reached obtained by controlling the VwCG levels individually. Through the threshold saturation phenomena at node X, much tighter threshold voltage

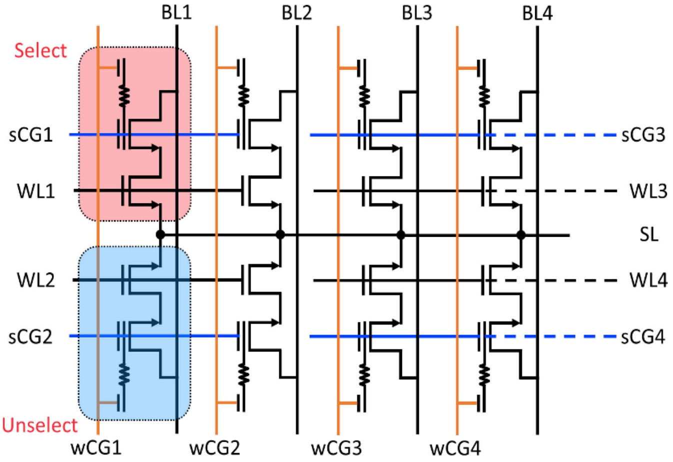


**FIGURE 7.** **The narrow-bridge FG cell arrayed in a NOR-type array, operated by block erase high and bit program low for multi-level storage.**

**TABLE 1.** **Summary of the operation conditions for the newly proposed MLC cell.**

| | | SCG | WCG | BL | WL | SL |
|---|---|---|---|---|---|---|
| Erase | | 12V | 0V | 0V | 0V | 0V |
| Program | Select | 0V | 10V/11V/12V(10μs) | 0V | 0V | 0V |
| | Unselect | 3.3V | 12V | 0V | 0V | 0V |
| Read | | 0V | 0V/1V/2V | 1.8V | 1.8V | 0V |

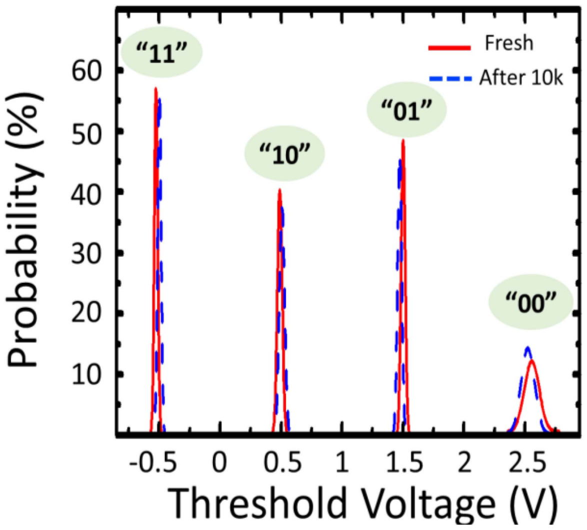


**FIGURE 8.** **Threshold voltage distribution between fresh and cycled cells with large sensing window maintained.**

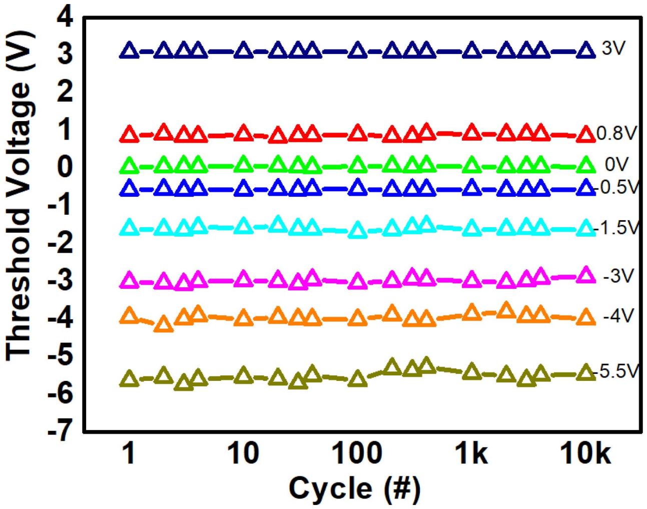


**FIGURE 9.** **TLC demonstration of 8 states maintaining good cyclability up to 10k.**

distributions can be attained without the need for precise control on the pulse width for programming.

Fig. 7 shows the cells arranged in a the NOR-type array, with block erase high and bit program low operation. With the sharing of wCG, a 3.3V are required to be applied on sCG to inhibit program disturb on unselected cell, preventing unwanted removal of written data. With this inhibit bias, a complete programming flow on the array level can be established. The disturb further characteristics indicate that a maximum 1K cells can share a common wCG without causing challenges to data integrity on the cells, suggesting a sizable array can be realized in this configuration. Table 1 is the operation table, which summarized the operation condition for MLC operation of the narrow-bridge FG cell.

The threshold voltage distributions before and after the 10k P/E cycling stress are compared in Fig. 8. Overall, through this self-limiting scheme and voltage control MLC operation, the multi-level Vth state of the narrow-bridge FG cells are much tightly controlled, enabling larger sensing windows. The 00 states are obtained by FN erase with applying a high positive PG to pull the electrons to FG from the channel of the FG transistor. During erase operation, the carrier injection goes through node Y, which subject to the charge accumulation effect causing clamping of $V_{TH}$ state. As shown in Fig. 8, the $V_{TH}$ spread of 00 state is less tighten. As a result of the self-limiting characteristic, the proposed operation also minimizes overstress during cycling, leading to tight Vth distributions even after 10k P/E cycles. To further expand the storage density of the proposed cell, a TLC operation under 10K endurance test is demonstrated in Fig. 9. Compare to normal FG structure, this newly proposed operation successfully prevents the threshold voltage levels from merging as a result of its self-limiting feature of the localizing charging effect.

## III. CONCLUSION

A novel narrow-bridge FG structure is demonstrated on the standard logic CMOS platform. By the self-limiting features, multi-level threshold voltage states can be reached without any extra circuit or complex programming algorithm. Finally, TLC operations has be demonstrated for high density data storage on this new cell.

## REFERENCES

[1] C.-Y. Lu, T.-C. Lu, and R. Liu, "Non-volatile memory technology-today and tomorrow," in *Proc. 13th Int. Symp. Phys. Failure Anal. Integr. Circuits*, Singapore, 2006, pp. 18–23.

[2] M. Bauer *et al.*, "A multilevel-cell 32 Mb flash memory," in *30th IEEE Int. Symp. Multiple Valued Logic Dig Tech. Papers (ISMVL)*, Portland, OR, USA, Feb. 1995, pp. 132–133.

[3] C. Matsui and K. Takeuchi, "Design of heterogeneously-integrated memory system with storage class memories and NAND flash memories," in *Proc. 24th Asia South Pac. Design Automat. Conf. (ASP-DAC UDC)*, 2019, pp. 17–18.

[4] Y.-H. Chang and T. W. Kuo, "A reliable MTD design for MLC flash-memory storage systems," in *Proc. 10th ACM Int. Conf. Embedded Softw. (EMSOFT)*, Scottsdale, AZ, USA, 2010, pp. 179–188.

[5] S. Lee *et al.*, "A 128 Gb 2b/cell NAND flash memory in 14nm technology with $t_{PROG}$=640$\mu$s and 800MB/s I/O rate," in *IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers (ISSCC)*, San Francisco, CA, USA, Jan./Feb. 2016, pp. 138–139.

[6] K. Kim, "Future memory technology: Challenges and opportunities," in *Proc. Int. Symp. VLSI Technol. Syst. Appl.*, Hsinchu, Taiwan, 2008, pp. 5–9.

[7] L. M. Grupp, J. D. Davis, and S. Swanson, "The bleak future of NAND flash memory," in *Proc. 10th USENIX Conf. File Storage Technol. (FAST12)*, Berkeley, CA, USA, 2012, p. 2.

[8] Q. Li, A. Jiang, and E. Haratsch, "Noise modeling and capacity analysis for NAND flash memories," in *Proc. IEEE Int. Symp. Inf. Theory*, Honolulu, HI, USA, Jun. 2014, pp. 2262–2266.

[9] J. Anxiao, R. Mateescu, M. Schwartz, and J. Bruck, "Rank modulation for flash memories," *IEEE Trans. Inf. Theory*, vol. 55, no. 6, pp. 2659–2673, Jun. 2009.

[10] A. Jiang, H. Li, and J. Bruck, "On the capacity and programming of flash memories," *IEEE Trans. Inf. Theory*, vol. 58, no. 3, pp. 1549–1564, Mar. 2012.

[11] A. Jiang and H. Li, "Optimized cell programming for flash memories," in *Proc. IEEE Pac. Rim Conf. Commun. Comput. Signal Process. (PACRIM)*, Victoria, BC, Canada, Aug. 2009, pp. 914–919.

[12] M. Qin, E. Yaakobi, and P. Siegel, "Optimized cell programming for flash memories with quantizers," *IEEE Trans. Inf. Theory*, vol. 60, no. 5, pp. 2780–2795, May 2014.

[13] W. Wang, T. Xie, A. Khoueir, and Y. Kim, "Reducing MLC flash memory retention errors through programming initial step only," in *Proc. IEEE 31st Symp. Mass Storage Syst. Technol. (MSST)*, Santa Clara, CA, USA, 2015, pp. 1–8.

[14] A. A. Chaudhry, C. Kui, and Y. L. Guan, "Mitigating stuck cell failures in MLC NAND flash memory via inferred erasure decoding," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 25, no. 8, pp. 2285–2295, Aug. 2017.

[15] Y. Kang and E. L. Miller, "Adding aggressive error correction to a high-performance compressing flash file system," in *Proc. 7th ACM Int. Conf. Embedded Softw. (EMSOFT)*, Grenoble, France, 2009, pp. 305–314.

[16] Y. Wang, L. A. D. Bathen, N. D. Dutt, and Z. Shao, "Meta-Cure: A reliability enhancement strategy for metadata in NAND flash memory storage systems," in *Proc. Design Automat. Conf. (DAC)*, San Francisco, CA, USA, 2012, pp. 214–219.

[17] R. Zhou, M. Liu, and T. Li, "Characterizing the efficiency of data deduplication for big data storage management," in *Proc. IEEE Int. Symp. Workload Characterization (IISWC)*, Portland, OR, USA, 2012, pp. 98–108.

[18] F. Sala, R. Gabrys, and L. Dolecek, "Dynamic threshold schemes for multi-level non-volatile memories," *IEEE Trans. Commun.*, vol. 61, no. 7, pp. 2624–2634, Jul. 2013.

[19] A. M. Caulfield, L. M. Grupp, and S. Swanson, "Gordon: Using flash memory to build fast, power-efficient clusters for data-intensive applications," in *Proc. 14th Int. Conf. Archit. Support Program. Lang. Oper. Syst. (ASPLOS'09)*, Mar. 2009, pp. 217–228.

[20] L. Chang, "A hybrid approach to NAND-flash-based solid-state disks," *IEEE Trans. Comput.*, vol. 59, no. 10, pp. 1337–1349, Oct. 2010.

[21] Y.-H. Yeh, C.-T. Chien, C. J. Lin, and Y.-C. King, "Embedded multiple-time programmable memory by BCD process for high voltage circuits," in *Proc. Int. Conf. Solid-State Devices Mater. (SSDM)*, Tokyo, Japan, Sep. 2018, pp. 1–2.

[22] D. Y. Wu, S. F. Chen, C. Lin, and Y. King, "Dummy read scheme for lifetime improvement of MLC NAND flash memories," *IEEE Trans. Device Mater. Rel.*, vol. 16, no. 4, pp. 583–587, Dec. 2016.

[23] K. C. Hsien, K. S. Chao, and Y. H. Chen, "Optimal Vth window in endurance and retention enhancement of MLC flash," in *Proc. 20th IEEE Int. Symp. Phys. Failure Anal. Integr. Circuits (IPFA)*, Suzhou, China, 2013, pp. 650–653.

[24] K.-T. Park, J. Choi, S. Cho, Y. Choi, and K. Kim, "A high cost-performance and reliable 3-level MLC NAND flash memory using virtual page cell architecture," in *Proc. 21st IEEE Non Volatile Semicond. Memory Workshop*, Monterey, CA, USA, 2006, pp. 34–35.

[25] M. Helm *et al.*, "19.1 A 128Gb MLC NAND-flash device using 16nm planar cell," in *IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers (ISSCC)*, San Francisco, CA, USA, 2014, pp. 326–327.

[26] L. Yao, D. Liu, K. Zhong, L. Long, and Z. Shao, "TLC-FTL: Workload-aware flash translation layer for TLC/SLC dual-mode flash memory in embedded systems," in *Proc. IEEE 17th Int. Conf. High Perform. Comput. Commun. 7th Int. Symp. Cybersp. Safety Security 12th Int. Conf. Embedded Softw. Syst.*, New York, NY, USA, 2015, pp. 831–834.

[27] A. Gehring and S. Selberherr, "Modeling of tunneling current and gate dielectric reliability for nonvolatile memory devices," *IEEE Trans. Device Mater. Rel.*, vol. 4, no. 3, pp. 306–319, Sep. 2004.