

Comprehensive SPICE Compact Modeling of Split-Gate Embedded Flash Memory for Transient and Worst-Case Operation Failure Simulation

Dongwoo Lim
Dept. of Electrical and Computer Eng.
Seoul National University
Seoul, South Korea
dw.lim@snu.ac.kr

Kiyoung Kim
Foundry Business
Samsung Electronics Co., Ltd.
Yongin-si, South Korea
kylsi.kim@samsung.com

Hyungcheol Shin
Dept. of Electrical and Computer Eng.
Seoul National University
Seoul, South Korea
hcshin@snu.ac.kr

Abstract—This paper presents a comprehensive SPICE compact modeling methodology for split-gate embedded flash (eFLASH) memory that supports transient program/erase simulation and worst-case operation-failure analysis. The effectiveness and accuracy of the proposed methodology are validated using experimental data from a 28-nm HKMG logic process. To accurately capture bias-dependent and time-transient threshold voltage (V_{TH}) shifts during program and erase operations, tunneling and injection current mechanisms are implemented in Verilog-A. Furthermore, cycling-induced endurance degradation and worst-case word-line (WL) transistor behavior are incorporated into a statistical modeling framework for program/erase failure analysis. The resulting model estimates high-sigma operation-failure probabilities within the SPICE environment, providing a practical basis for reliability-aware eFLASH design.

Index Terms—eFLASH, Split-Gate, SPICE Modeling, Statistical Modeling, Endurance Modeling, Reliability Simulation

I. INTRODUCTION

As embedded flash (eFLASH) is increasingly deployed in reliability-critical applications, including AI, automotive, IoT, and security [1, 2], accurate SPICE modeling has become indispensable for ensuring design robustness. The conventional BSIM-based SPICE model [3] is widely adopted in industry due to its simplicity in predicting read current (I_{READ}); however, its reliance on empirical equations for floating-gate charge (Q_{FG}) limits the modeling to DC bias-dependent behavior and fails to capture time-transient Q_{FG} evolution during program and erase operations. In addition, such models are typically calibrated to nominal process corners, making it difficult to adequately cover worst-case scenarios.

To address these limitations, this work presents a unified modeling framework with two key components. First, compact modeling of Fowler–Nordheim (FN) tunneling and Source-Side Injection (SSI) currents enables the dynamic prediction of Q_{FG} as a function of time and bias, from which quantitative threshold voltage (V_{TH}) shifts are derived. Second, large-volume measurement data are used to construct statistical models that capture cycling-induced endurance degradation and worst-case WL transistor behavior. This framework enables SPICE-based prediction of erase and program failure

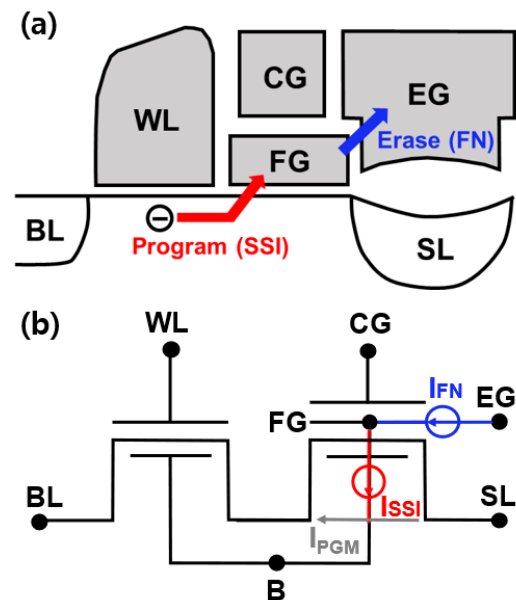


Fig. 1. Split-gate embedded flash cell structure. (a) Cross-sectional schematic of the split-gate eFLASH cell with electron-flow paths during program and erase operations. (b) Corresponding SPICE equivalent circuit model.

cases as well as the estimation of high-sigma failure levels, facilitating more robust and reliability-aware design planning.

II. MODELING FOR TRANSIENT SIMULATION

This section presents the compact modeling procedure for transient program and erase simulation, along with the corresponding simulation results.

A. DC Modeling and Coupling Ratio Extraction

The DC I–V characteristics are first modeled following the methodology in [3], because Q_{FG} -induced ΔV_{TH} is ultimately reflected in I_{READ} through the read-current model. As shown in Fig. 2(a) and (b), the measured DC I_{BL} characteristics are fitted as functions of V_{WL} and V_{BL} , respectively, confirming

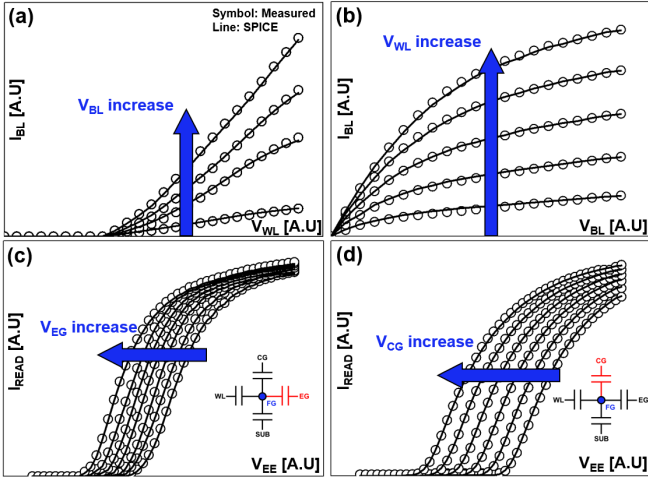


Fig. 2. DC I-V fitting and coupling-ratio extraction based on the methodology in [3]. (a) and (b) show the I_{BL} characteristics as functions of V_{WL} and V_{BL} , respectively. (c) and (d) show the node-specific coupling ratios extracted by fitting the node-voltage dependence of the erase curve.

$V_{FG}(t) = CR_{CG} \cdot V_{CG} + CR_{EG} \cdot V_{EG} + CR_{WL} \cdot V_{WL} + CR_{SUB} \cdot V_B + \Delta Q_{FG}(t)/C_{eff}$		
$I_{FN} = A \cdot E_{FN}^2 \exp\left(-\frac{B}{E_{FN}}\right)$ $A = A_{FN} \cdot \frac{q^3}{16\pi\hbar\phi_B} \quad B = B_{FN} \cdot \frac{4\sqrt{2m_{ox}^*}}{3q\hbar} \phi_B^{3/2}$	FN Tunneling	$\Delta Q_{FG} = \int_0^t I_{FN,SSI} dt$ feedback
$I_{SSI} = I_{PGM} \cdot p_{SSI}$ $p_{SSI} = A_{inj} \cdot \exp\left[-B_{inj} \left(\frac{\phi_B}{kT_e}\right)^n\right] \cdot \exp\left[-C_{inj} \left(\frac{1}{E_{ox}}\right)\right]$	SSI (Lucky Electron Model)	
$\phi_B[\text{eV}] = 3.1 - \beta E_{ox}^{1/2} - \alpha E_{ox}^{2/3}$	[Fitting Parameters]	
$A_{FN}, B_{FN}, A_{inj}, B_{inj}, C_{inj}, V_{FG}(0)$		
$\Delta V_{TH}(t) = V_{FG}(t) - V_{FG}(0) = \Delta Q_{FG}(t)/C_{eff} \rightarrow \text{feedback to } V_{CG}$		

Fig. 3. Compact model equations for transient Q_{FG} and I_{READ} simulation. The calculated ΔV_{TH} is applied as an equivalent offset to V_{CG} , and the newly introduced C_{inj} term captures the saturation behavior of the program curve in Fig. 4(b).

that the model reproduces the bias-dependent read-current behavior. In addition to the curves shown in Fig. 2, the DC fitting was performed over all bias and temperature conditions considered in this work.

The node-specific coupling ratios (CRs) are then extracted by fitting the node-voltage dependence of the erase curve using the corresponding capacitance values, as summarized in Fig. 2(c) and (d). The extracted CRs are subsequently incorporated into the transient Q_{FG} model to calculate the Q_{FG} -induced threshold voltage shift (ΔV_{TH}) together with the effective capacitance (C_{eff}), which is experimentally extracted as described in [4].

B. Floating Gate Charge Modeling

Erase and program operations are modeled using the I_{FN} and I_{SSI} current sources shown in Fig. 1(b). The corresponding model equations are summarized in Fig. 3 [5, 6]. The calculated ΔV_{TH} is applied as an equivalent voltage offset to the

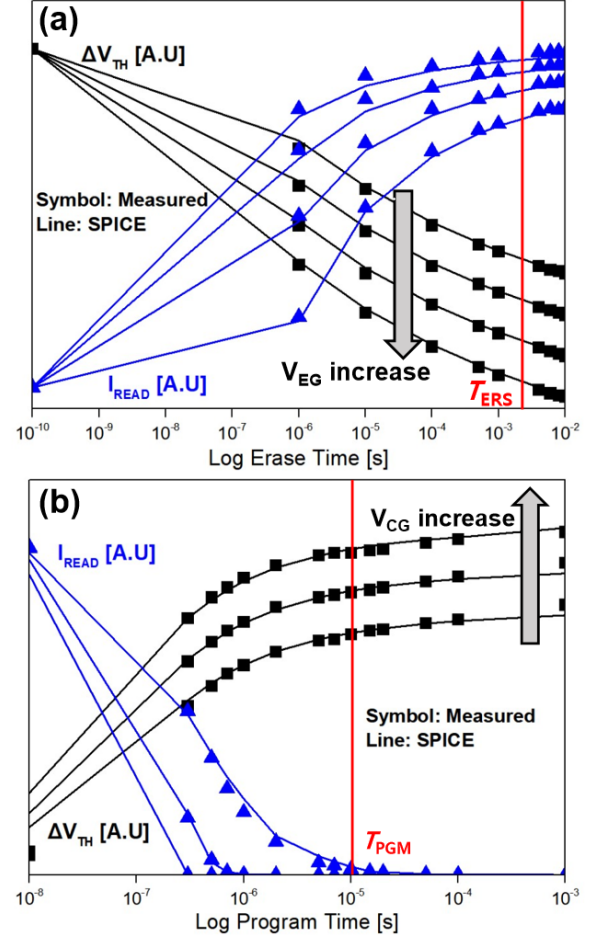


Fig. 4. Comparison of measured and simulated ΔV_{TH} and I_{READ} as functions of (a) erase time under 1-V-stepped V_{EG} conditions and (b) program time under 1-V-stepped V_{CG} conditions.

control-gate voltage (V_{CG}) under read conditions, allowing the threshold shift to be reflected in the transient I_{READ} simulation. Fig. 4(a) and (b) compare the SPICE simulation results with measured data, demonstrating accurate prediction of transient behavior over a wide range of time and bias conditions.

In the Verilog-A implementation, the transient Q_{FG} dynamics are implemented using an external node driven by the operation-bias voltages defined in the SPICE netlist. The transient block is coupled to the conventional DC-based read model only through equivalent voltage feedback. Since the model computes ΔV_{TH} through compact equations, the run-time overhead of transient simulation is minimal.

III. WORST-CASE MODELING

This section describes three worst-case factors incorporated into the proposed SPICE framework: statistical process variation, WL-transistor current limitation, and cycling-induced endurance degradation.

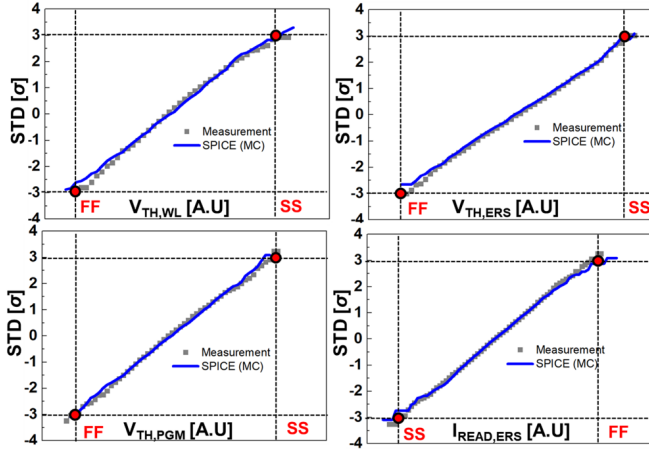


Fig. 5. Comparison of large-volume experimental data with Monte Carlo and fixed-corner modeling results.

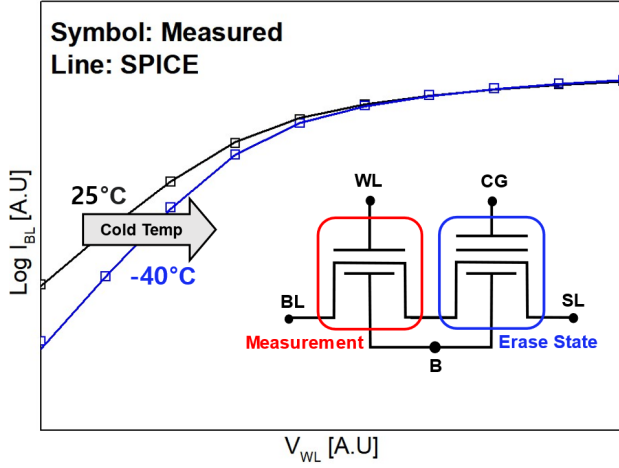


Fig. 6. Measured and modeled I-V characteristics of the WL transistor at 25 °C and -40 °C. The measurements were performed in the erased state to isolate the WL transistor behavior from the cell-transistor characteristics.

A. Statistical Modeling

As eFLASH-embedded logic technologies scale down, the impact of process variations becomes increasingly pronounced. Local and global variation components are characterized separately: in-die variations are treated as local components, whereas the remaining variation sources are modeled as global components. The resulting total variation is compared with large-volume experimental data in Fig. 5, along with the Monte Carlo and fixed-corner modeling results. The backward propagation of variation (BPV) methodology [7] is employed. In this approach, BSIM parameters (TOX, XL, NFACTOR, etc.) as well as $Q_{FG,PGM}$ and $Q_{FG,ERS}$ are defined using Gaussian distributions.

B. Word-Line Transistor Modeling

Accurate I-V modeling of the word-line (WL) transistor has also become essential. While earlier studies often treated the WL transistor as a simple switch, continued scaling has made

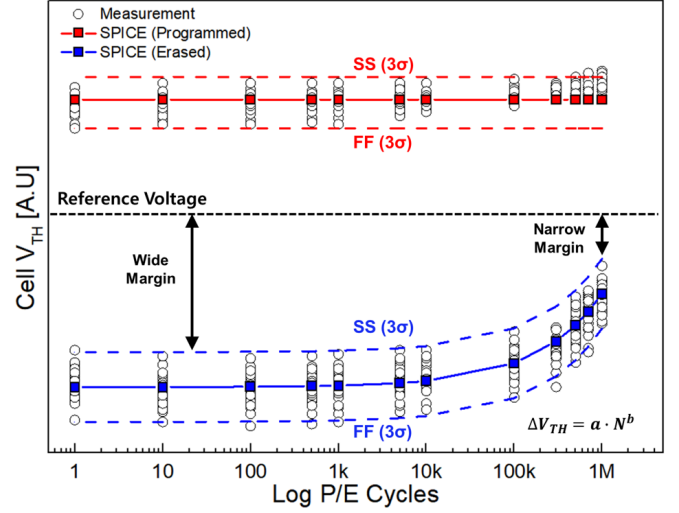


Fig. 7. Measured and modeled programmed and erased V_{TH} as a function of program/erase cycling count (N). The erased-state degradation is fitted using a power-law model.

device mismatch and temperature dependence non-negligible. In particular, device scaling and reduced operating voltage decrease the WL overdrive margin; at low temperature, the WL transistor V_{TH} can approach the applied V_{WL} , limiting I_{PGM} due to insufficient turn-on. As shown in Fig. 6, the intrinsic I-V characteristics of the WL transistor are extracted with the cell transistor in the erased state and are modeled using BSIM parameters.

C. Endurance Modeling

In split-gate architectures with a separate erase gate (EG), cycling-induced degradation is observed predominantly in the erased-state V_{TH} [8]. Because this degradation can result in the worst-case read-fail condition, the model incorporates only the erased-state V_{TH} degradation. Fig. 7 presents the measured programmed and erased V_{TH} as a function of cycling count (N), along with the modeling results fitted to a power-law equation [9]. The cycling count N can be defined as an instance parameter, and the calculated ΔV_{TH} can be reflected in the BSIM parameters of the cell transistor.

IV. HIGH-SIGMA OPERATION FAILURE ANALYSIS

By incorporating the worst-case factors modeled in Section III, cell-level program and erase failure probabilities can be predicted within the SPICE framework. Although modeling is typically performed at the 3σ level based on measurement data, extrapolation beyond the measured range is a commonly adopted methodology for evaluating high-sigma behavior in high-reliability applications [10].

Erase operation failure may occur in the high-sigma tail under a large cycling count (N). As illustrated in Fig. 8(a), after 1M program/erase cycles, the proposed approach enables prediction of the sigma levels at which I_{READ} falls below the sensing criterion.

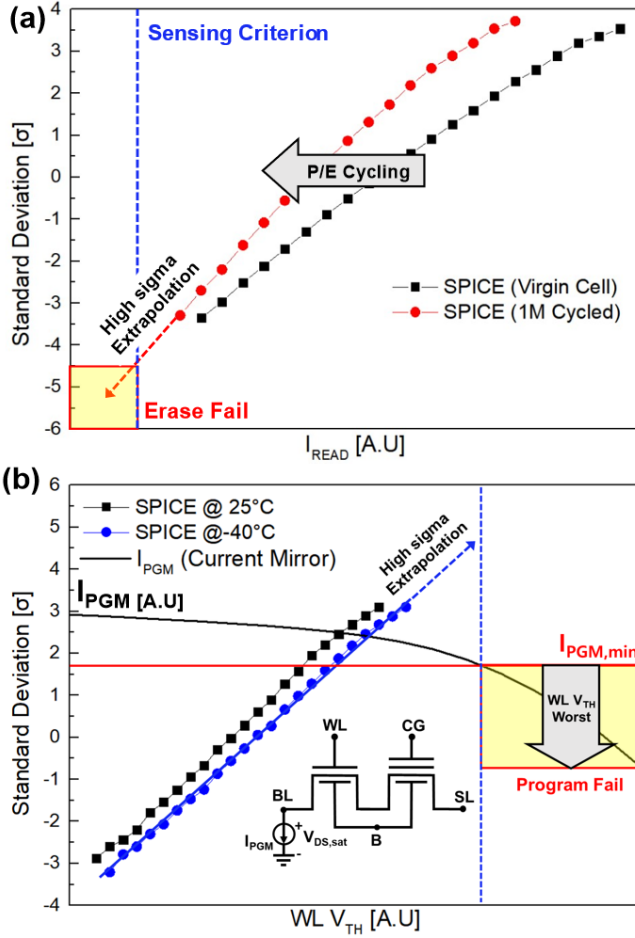


Fig. 8. High-sigma failure analysis for (a) erase operation (I_{READ} below the sensing criterion) and (b) program operation (I_{PGM} below the functional limit), considering cycling-induced read-margin reduction and worst-case WL transistor behavior. The minimum bit-line voltage (V_{BL}) in (b) is determined by the current-mirror $V_{\text{DS,sat}}$.

Program operation failure can occur when the WL transistor cannot supply sufficient current, causing I_{PGM} to fall below the functional limit. In this evaluation, the I_{PGM} source is designed as a two-stage cascode current mirror, and the simulation is performed at -40°C . As shown in Fig. 8(b), the proposed approach estimates the high-sigma conditions under which the WL transistor fails to supply the required I_{PGM} .

V. CONCLUSION

We proposed a SPICE compact modeling methodology for split-gate eFLASH cells that supports transient program/erase simulation and worst-case operation-failure analysis. FN tunneling and SSI current models captured transient Q_{FG} behavior and accurately reproduced ΔV_{TH} and I_{READ} under various program/erase biases. Process variation, endurance degradation, and low-temperature WL-current limitation were incorporated as worst-case factors. This approach enables SPICE-based evaluation of erase failure due to read-margin loss and program failure due to insufficient I_{PGM} .

REFERENCES

- [1] N. Do, H. Tran, and M. Reiten, "Computing-in-Memory with SuperFlash[®] memBrain[™] Technology," in *Proceedings of the IEEE Symposium on VLSI Technology and Circuits*, 2022, pp. 238–239.
- [2] Y. K. Lee *et al.*, "High-Speed and Logic-Compatible Split-Gate Embedded Flash on 28-nm Low-Power HKMG Logic Process," in *Proceedings of the IEEE Symposium on VLSI Technology*, 2017, pp. T202–T203.
- [3] S. Martinie, O. Rozeau, M. Tadayoni, C. Raynaud, E. Nowak, S. Hariharan, and N. Do, "SPICE Modeling of 55 nm Embedded SuperFlash[®] Technology 2T Memory Cells," in *Proceedings of the IEEE International Conference on Microelectronic Test Structures (ICMTS)*, 2015, pp. 24–28.
- [4] Y. Tkachev and A. Kotov, "Detection of Single-Electron Transfer Events and Capacitance Measurements in Submicron Floating-Gate Memory Cells," in *Proceedings of the European Solid-State Device Research Conference (ESSDERC)*, 2006, pp. 411–414.
- [5] G. J. Hemink, R. C. M. Wijburg, P. B. M. Wolbert, and H. Wallinga, "Modeling of VIPMOS Hot Electron Gate Currents," *Microelectronic Engineering*, vol. 15, no. 1–4, pp. 65–68, 1991.
- [6] Y.-H. Wang, M.-C. Wu, C.-J. Lin, W.-T. Chu, Y.-T. Lin, C. S. Wang, and K.-Y. Cheng, "An Analytical Programming Model for the Drain Coupling Source Side Injection Split Gate Flash EEPROM," *IEEE Transactions on Electron Devices*, vol. 52, no. 3, pp. 385–391, Mar. 2005.
- [7] X. Li, C. C. McAndrew, W. Wu, S. Chaudhry, J. Victory, and G. Gildenblat, "Statistical Modeling With the PSP MOSFET Model," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 29, no. 4, pp. 599–606, Apr. 2010.
- [8] Y. Tkachev and A. Kotov, "A New Method for Analysis of Cycling-Induced Degradation Components in Split-Gate Flash Memory Cells," in *Proceedings of the IEEE International Reliability Workshop (IIRW)*, 2013, pp. 117–120.
- [9] Y. Luo, S. Ghose, Y. Cai, E. F. Haratsch, and O. Mutlu, "Enabling Accurate and Practical Online Flash Channel Modeling for Modern MLC NAND Flash Memory," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 9, pp. 2294–2311, Sep. 2016.
- [10] H.-H. Wang, Y.-L. Chen, C.-C. Yang, C.-K. Lin, and M.-C. Jeng, "Compact Modeling for Application-Specific High-Sigma Worst Case," in *Proceedings of the International Conference on Simulation of Semiconductor Processes and Devices (SISPAD)*, 2013, pp. 61–64.