

Simulation Study on Scaling Limits of Z-direction Pitch in 3D Flash Memories: Toward the Sub-40 nm Regime

Tatsuo Ogura, Yuma Hizume, Yuto Ozeki, Takayuki Kakegawa, Hiroshi Takeda, Takashi Kurusu, Tomohiro Oki, and Katsuyuki Sekine

Advanced Memory Development Center, Kioxia Corporation, Yokkaichi, Mie, Japan

e-mail: tatsuo.ogura@kioxia.com

Abstract—Three-dimensional (3D) flash memories achieve high-density storage by stacking multiple layers; however, scaling down the word-line (WL) pitch increases cell-to-cell interference, degrades data retention, and enhances random telegraph noise (RTN). Breakthrough techniques—metal-induced lateral crystallization (MILC) and charge-trap layer segmentation (CTS)—address these issues by suppressing lateral charge diffusion and mitigating the adverse effects of grain boundaries in silicon channels. In this work, we develop a multiscale simulation framework combining TCAD, empirical models, and Monte Carlo methods to predict the scaling limits of the WL pitch. Conventional cells struggle to operate below a WL pitch of 40 nm for quad-level cells (QLC), whereas the combined use of CTS and MILC enables sub-40 nm scaling and makes penta-level cells (PLC) feasible near the 40 nm node. These results provide insight into overcoming scaling challenges in advanced 3D flash memories.

Keywords—3-dimensional, Flash Memory, Simulation, threshold voltage variability

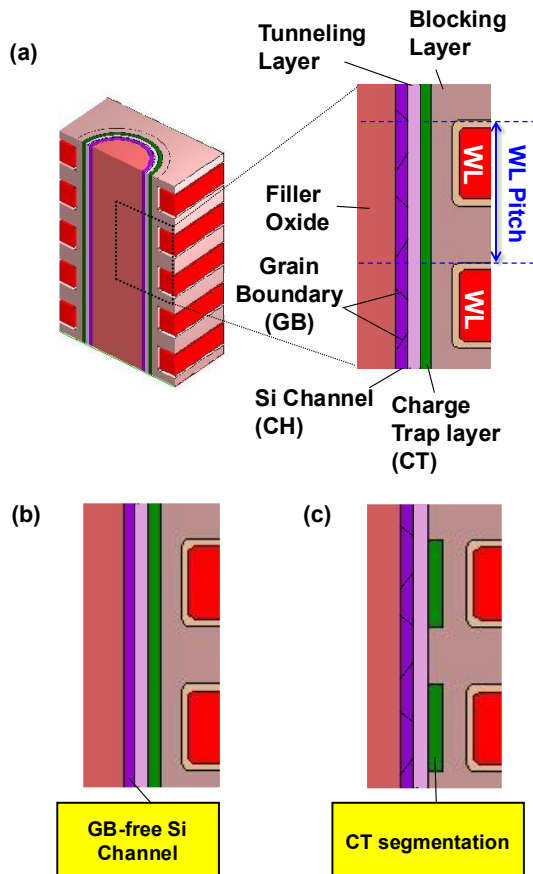


Fig. 1. (a) Bird's-eye view of a conventional memory cell array and its cross-sectional views with key feature dimensions; (b) cell structure with MILC; and (c) cell structure with CTS.

I. INTRODUCTION

With the growing demand for higher density and better performance in non-volatile memory technologies, three-dimensional (3D) flash memories have been developed by vertically stacking multiple layers [1]. However, increasing the number of layers and scaling down cell dimensions introduce significant physical and electrical challenges that affect device performance and reliability. Scaling the word-line (WL) pitch along the Z-direction increases cell-to-cell (C2C) interference [2] and lateral charge migration [3], thereby degrading performance and retention characteristics. In addition, WL pitch scaling increases the sensitivity of the threshold voltage (V_{th}) to trapped charges, thereby enhancing random telegraph noise (RTN), which is primarily caused by trap states at grain boundaries (GBs) in the polysilicon channel [4]. Ultimately, these challenges have brought Z-direction scaling to its limits, thereby hindering further cost reductions. Two breakthrough techniques—metal-induced lateral crystallization (MILC) [5] and charge-trap layer segmentation (CTS) [6]—have been proposed to

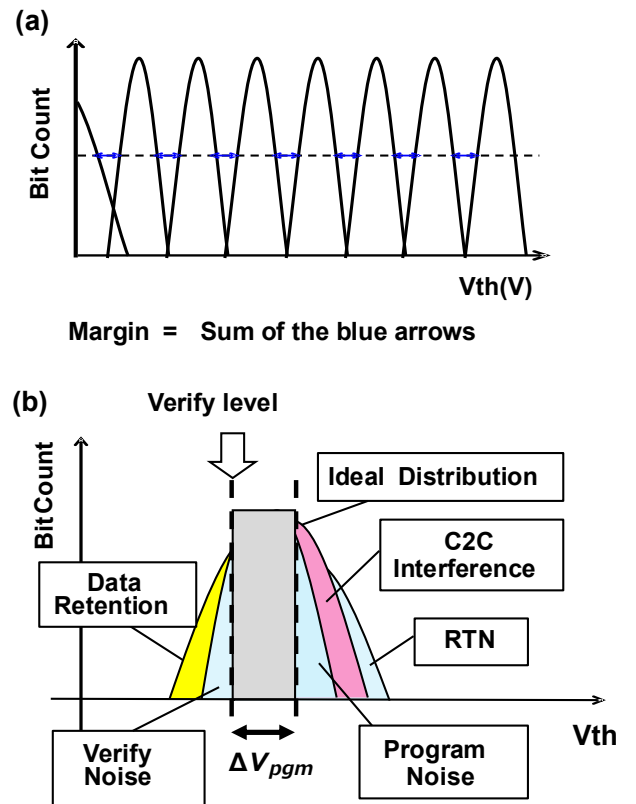


Fig. 2. (a) Definition of the V_{th} budget in a multi-level cell and (b) key factors contributing to V_{th} margin degradation.

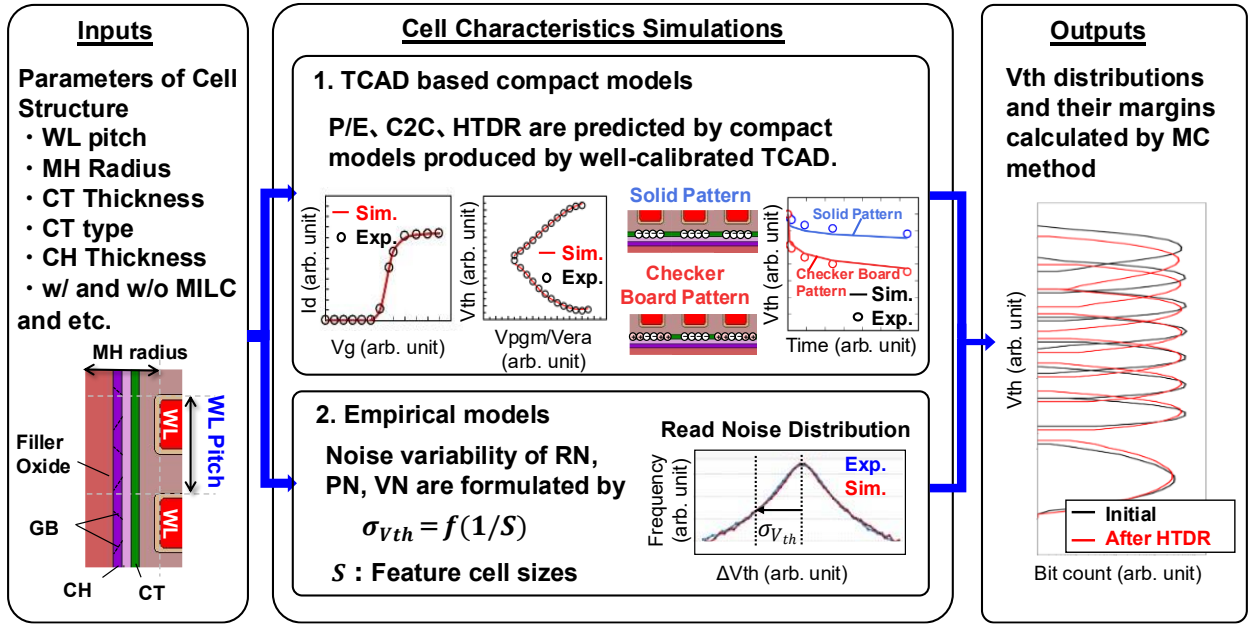


Fig. 3 Schematic of the simulation framework used in this study.

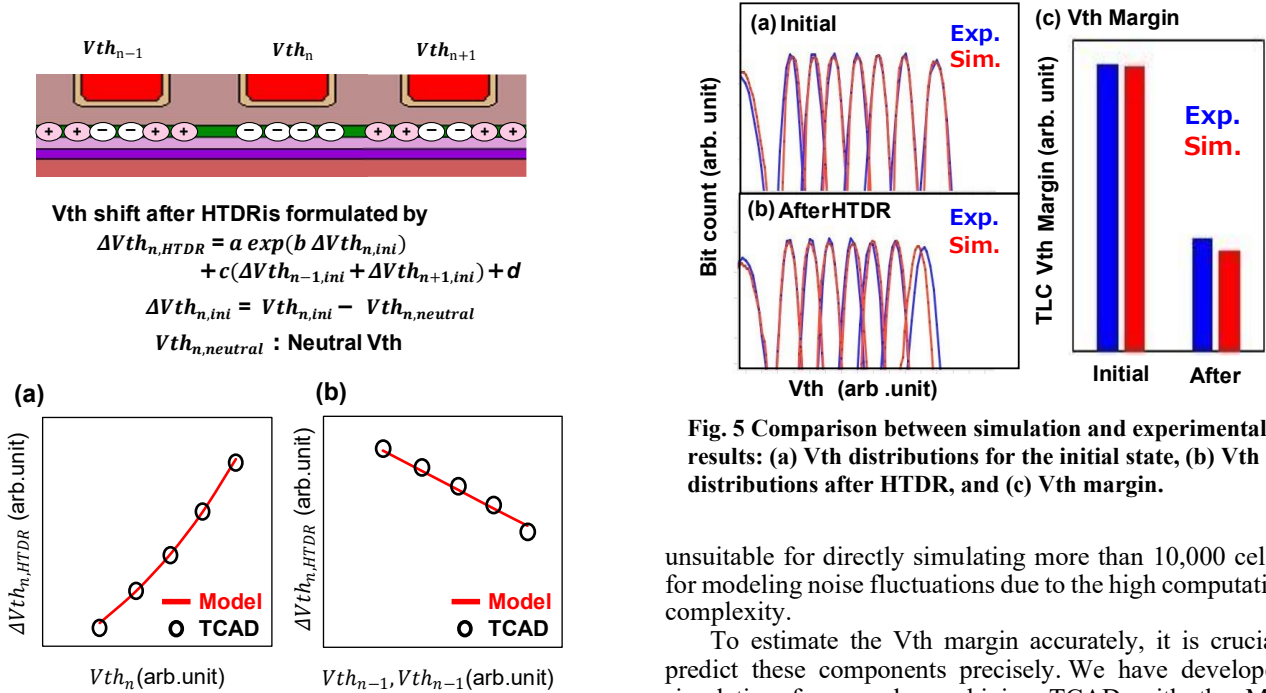


Fig. 4 Compact model for Vth shift after thermal baking, showing its dependence on (a) the target cell Vth and (b) the adjacent cell Vth.

overcome these limitations by reducing trap sites in the silicon channel and eliminating lateral diffusion of stored charge in the charge-trap layer (CT) (Fig. 1).

The Vth distribution in multi-level cell arrays, where each word line (WL) contains thousands of cells, is governed by program/erase (P/E) characteristics and high-temperature data retention (HTDR). The distribution is also affected by various noise sources, such as verify noise (VN), random telegraph noise (RTN), and program noise (PN). Specifically, verify noise (VN) refers to fluctuations in Vth that occur when the cell state changes between verify and read operations (Fig. 2) [7]. However, TCAD simulations are

unsuitable for directly simulating more than 10,000 cells or for modeling noise fluctuations due to the high computational complexity.

To estimate the Vth margin accurately, it is crucial to predict these components precisely. We have developed a simulation framework combining TCAD with the Monte Carlo (MC) method [8] to calculate the Vth distribution and determine the scaling limits of Z-direction pitch in 3D flash memories employing the MILC and CTS techniques.

II. SIMULATION METHODOLOGY

Figure 3 shows the simulation framework used in this study. Compact models of cell characteristics were developed using well-calibrated TCAD simulations and experimental data, and coupled with the MC method to calculate Vth distributions and margins. This framework enables efficient and accurate prediction of Vth distributions and margins.

Parameters for mobility and trap models in TCAD, for both the Si channel and the CT, were calibrated using prototype devices: mobility parameters were fitted to current-voltage (I-V) curves, and trap parameters were fitted to P/E curves

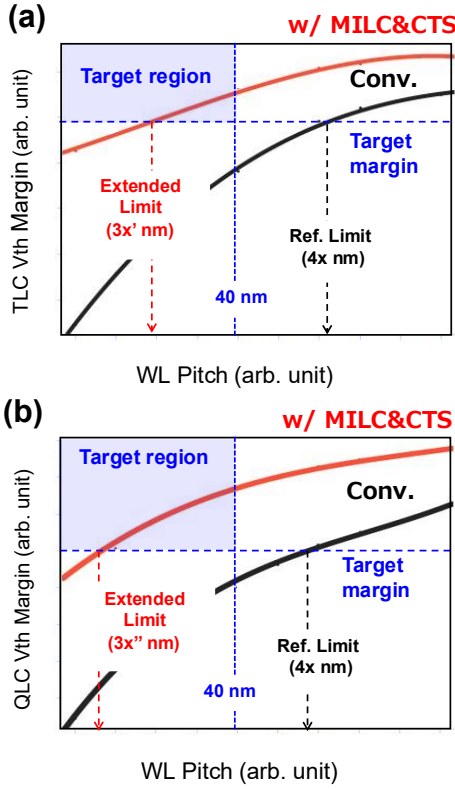


Fig. 6. Dependence of Vth margin on WL pitch for (a) TLC and (b) QLC. The target regions indicate areas that satisfy both the target Vth margin and a WL pitch below 40 nm.

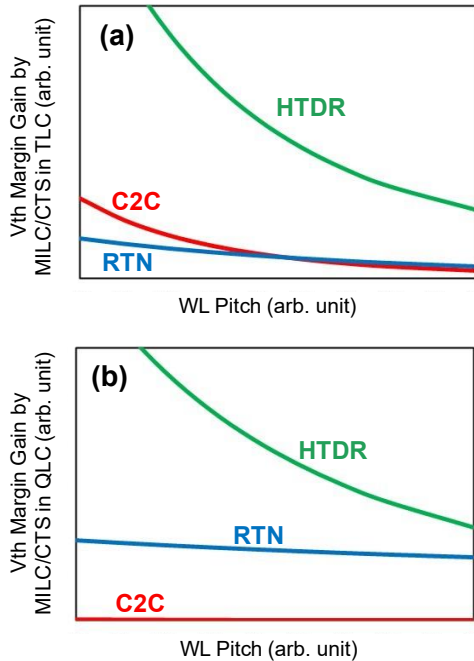


Fig. 7 Contributions to Vth width improvement for (a) TLC (3 bits per cell) and (b) QLC (4 bits per cell) with MILC and CTS.

and HTDR characteristics. Empirical models for RTN, VN, and PN, expressed as functions of cell feature size, were also calibrated using experimental data from the prototype devices.

The MC method can predict not only the initial Vth distribution but also the Vth distribution after HTDR, based on a compact model for Vth shift after high-temperature baking (thermal stress), expressed as a function of the Vth of the target cell and that of the adjacent cell (Fig. 4). The simulation framework was systematically confirmed to efficiently and quantitatively reproduce the Vth margins of real devices, both for fresh (as-programmed) cells and after HTDR (Fig. 5). The Vth margin decreases after HTDR due to a downward shift in Vth and broadening of the distribution. Using the Vth margin after HTDR as a key metric, we investigated the scaling limits of the WL pitch in 3D flash memories.

III. RESULTS AND DISCUSSION

Figure 6 shows the simulated dependence of the Vth margin after HTDR on the WL pitch for triple-level cells (TLC) and quad-level cells (QLC), both with and without MILC and CTS. In these simulations, a conventional programming scheme and a multi-stage programming scheme [7] were assumed for TLC and QLC, respectively. The results show that the Vth margin decreases as the WL pitch is scaled down and approaches its limit at around 40 nm. In contrast, the combination of MILC and CTS can effectively provide an approximately 10 nm improvement in scaling. Our study also reveals that C2C interference, HTDR, and RTN are the key factors limiting TLC scaling (Fig. 7a), whereas HTDR and RTN dominate in QLC (Fig. 7b). These differences arise from the programming schemes and target Vth margins. C2C interference and HTDR are mitigated by CTS, which reduces electrostatic coupling to adjacent cells and suppresses lateral diffusion of trapped carriers during HTDR (Fig. 8). RTN is reduced by the GB-free Si channel enabled by MILC.

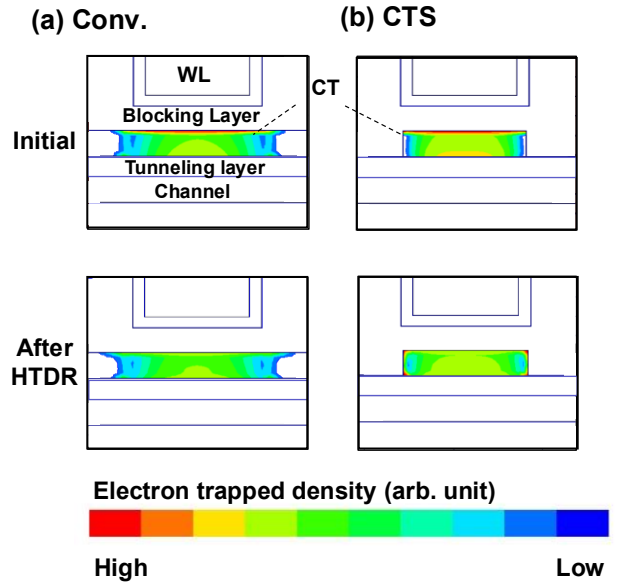


Fig. 8 Trapped electron distribution for (a) Conv. cells and (b) CTS cells.

Finally, the scaling limits of penta-level cells (PLC) (Figs. 9 and 10) were evaluated, which are currently in the research and development stage. The simulation results indicate the potential for PLC operation at the 4x nm node with these breakthrough technologies. Further work will focus on exploring optimal structures and developing new programming schemes to extend PLC scaling limits.

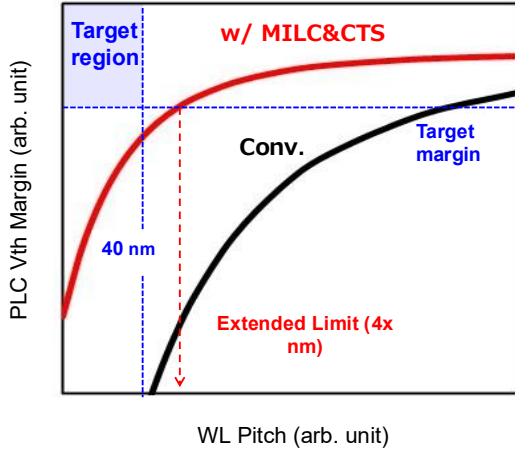


Fig. 9 Vth margin of PLC as a function of WL pitch.

IV. CONCLUSIONS

Using our efficient and accurate simulation framework combining TCAD with the MC simulations, it is demonstrated that conventional TLC and QLC cells cannot achieve a WL pitch below 40 nm without introducing new techniques. Cells incorporating both CTS and MILC enable scaling to a sub-40-nm WL pitch. However, the scaling limit for PLC remains around the 4x nm range even when MILC and CTS are applied. Therefore, further technological innovations are required to enable continued scaling of PLC.

ACKNOWLEDGMENT

This work was supported by the joint development of 3D Flash Memory with Sandisk Corporation.

REFERENCES

- [1] H. Tanaka *et al.*, *Proc. IEEE Symp. VLSI Technol.*, 2007, pp. 14–15.
- [2] H. Jo *et al.*, *IEEE Trans. Electron Devices*, vol. 69, 2, pp. 543–548, 2022.
- [3] D. Oh *et al.*, *Proc. IEEE Int. Memory Workshop.*, pp.1–4, 2015.
- [4] M.-K. Jeong *et al.*, *Proc. Symp. VLSI Technol.*, pp. 55–56, 2012.
- [5] N. Ishihara *et al.*, *Proc. IEEE Symp. VLSI Technol.*, 2023, pp. 14–15.
- [6] S.-W. Jin *et al.*, *Tech Digest of IEDM*, Dec. 2025, 18-3.J.
- [7] Lee *et al.*, *Solid State Electronics*, 229, 109198, 2025
- [8] H. Li, *IEEE Trans. Electron Devices*, vol. 63, no. 9, 2016, pp. 3527

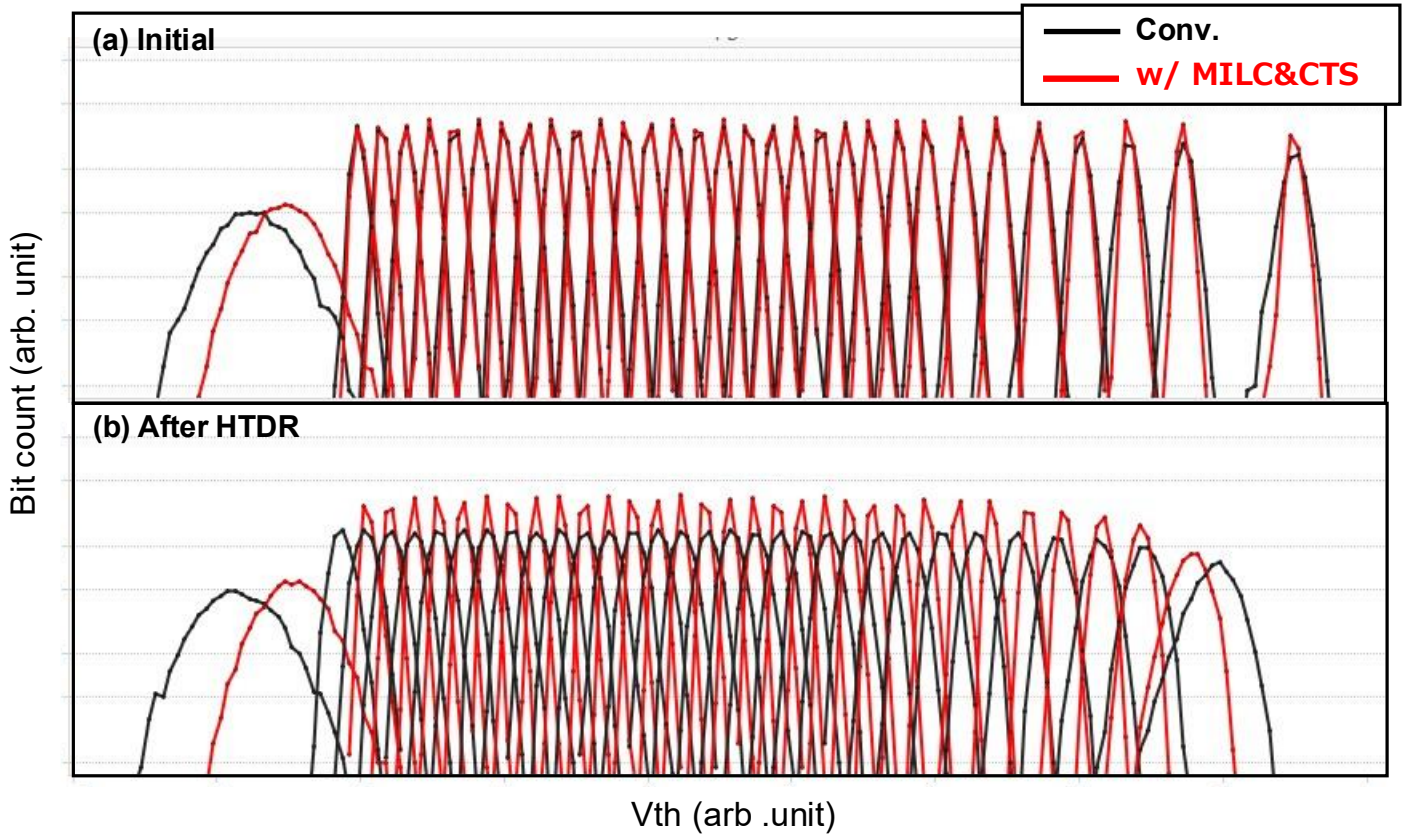


Fig. 10 Vth distribution of Conv. and MILC/CTS cells for PLC. (a)Initial and (b) after HTDR.