A Provably Optimized Approach to Predict Bit Error Rate Growth During Room Temperature Data Retention in Advanced 3-D NAND Flash Memories

Y. H. Liu[1], Y. S. Yang[1], C. C. Hsu[1], Y. C. Hsu[1], and W. Lin[1,2]
[1]Phison Electronics Corp., Miaoli, Taiwan
[2] College of Artificial Intelligence, National Yang Ming Chiao Tung University, Hsinchu, Taiwan
e-mail: henry_liu@phison.com

***Abstract*—Given that activation energy (E_a) varies with temperature, the conventional temperature-accelerated life-test methodology is no longer adequate for evaluating the reliability of advanced 3-dimensional (3-D) NAND Flash memories. Consequently, the ability to accurately estimate device lifetime based on limited room-temperature data retention measurements has become essential for modern reliability assessment. To address this challenge, we present a comparative analysis of three distinct prediction models: linear-linear extrapolation, log-linear extrapolation, and log-log extrapolation. To rigorously determine which model offers the highest predictive capability, we collected one year of room-temperature data retention results from eight different 3-D NAND Flash types and performed extensive validation using physics-based modeling and large-scale datasets. The results confirm that log-log extrapolation model consistently delivers superior predictive accuracy for room-temperature data retention, making it the most effective approach among the models evaluated.**

Keywords- 3-D NAND Flash memory, life-test methodology, log-log extrapolation, room-temperature data retention

I. INTRODUCTION

Driven by growing demand for enterprise cloud storage and artificial intelligence computing, the market for 3-D charge-trap (CT) NAND Flash memory has expanded significantly due to its ultrahigh density and low bit cost [1]–[3]. To further enhance the storage capacity of 3-D NAND Flash, scaling down both the cell channel length and the inter-gate spacing (IGS) within a NAND string is essential. However, such dimensional scaling introduces new reliability challenges. In particular, the reduction in cell size and the IGS region exacerbates lateral charge migration within the silicon nitride (SiN) charge-trap layer during data retention [4], [5], in addition to the conventional vertical charge loss through the tunnel oxide [6]–[8]. Our previous study [9] demonstrated that the conventionally employed high-temperature-accelerated methodology based on the Arrhenius equation [10], [11]—which assumes a single dominant charge loss mechanism—is no longer applicable. Figure 1 shows the extracted characteristic retention times (τ) required to reach a specified criteria of V_t shift at various bake temperatures. These τ values are plotted as a function of the inverse absolute temperature in the inset of Fig. 1. The variation in dominant physical mechanisms across different temperature regimes results in a temperature-dependent E_a, which ultimately causes an overestimation of data retention capability when using simplified models. In this work, we also collected data across 3-D NAND Flash products from multiple vendors to extract

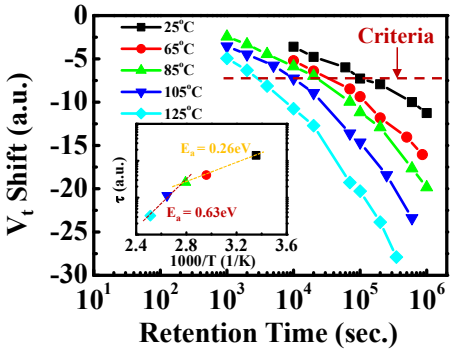

Fig. 1. Evolutions of V_t shift with retention time at various bake temperatures. Inset: A certain retention time (τ) to a criteria of V_t shift at various bake temperatures (T = 25°C, 65°C, 85°C, 105°C, and 125°C).

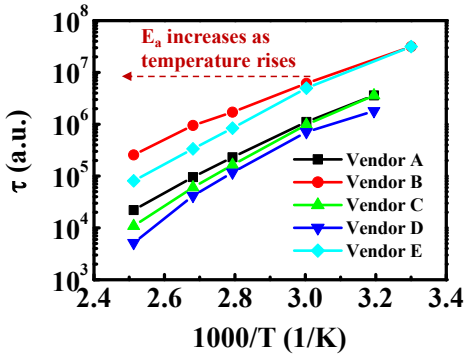

Fig. 2. A certain retention time (τ) to a criteria of raw bit error rate (RBER) at various bake temperatures (T = 25°C, 65°C, 85°C, 105°C, and 125°C) among 5 distinct types of TLC NAND.

the characteristic retention time (τ) required to reach a certain raw bit error rate (RBER), and illustrated how τ varies with temperature. It should be noted that the trend of E_a increasing with temperature is a universal behavior consistently observed among NAND Flash devices, as indicated in Fig. 2.

To mitigate the reliability assessment inaccuracies resulting from the divergence of E_a at different temperatures, the most accurate method for evaluating retention lifetime is to conduct long-term testing under the product's actual operating temperatures. However, performing real-time retention tests spanning a year or longer for each unit prior to shipment is impractical. Consequently, developing a reliable room-temperature data retention prediction model with broad

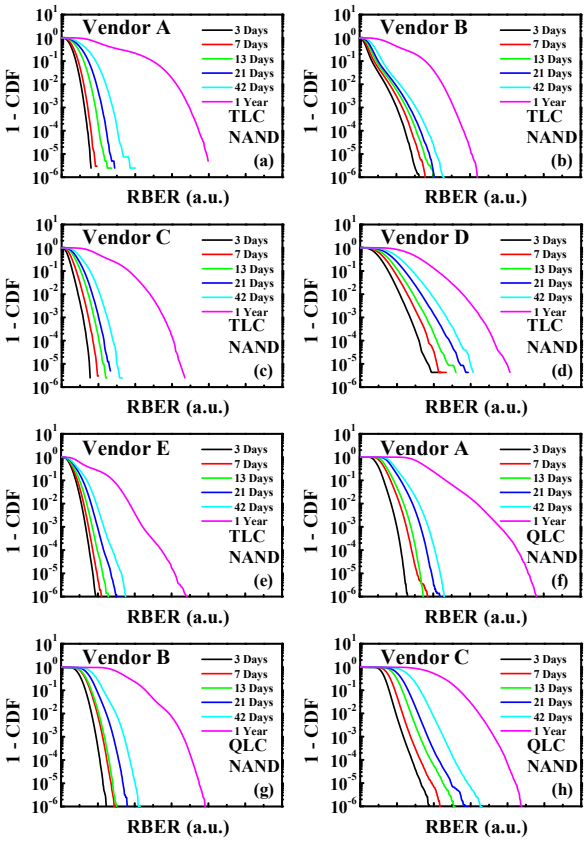

Fig. 3. Counter cumulative distribution (1-CDF) of RBER in retention at P/E cycle = 1 with 8 different types (5 TLC and 3 QLC) of NAND Flash. Retention T = 30°C or 40°C.

applicability across various NAND Flash brands is essential to enable accelerated testing and ensure product reliability within a feasible timeframe. Although various retention prediction models have been developed for planar multi-level-cell (MLC) NAND Flash over the past decade [12]–[14], and recently for 3-D triple-level-cell (TLC) NAND [15]–[17], critical gaps persist. Specifically, the method proposed in [15] lacks empirical validation, while the approach in [16] was calibrated against a single NAND vendor, undermining its generalizability. Conversely, although [17] incorporated empirical data from three vendors, the resulting model suffers from an excessive number of fitting parameters. To address these limitations, this study comprehensively evaluates and compares three distinct prediction frameworks: linear-linear, log-linear, and log-log extrapolation models. To verify their accuracy, a one-year data retention experiment was conducted across eight distinct types of commercial NAND Flash chips. Furthermore, the physical validity of these models is rigorously substantiated through their underlying mechanisms.

II. NAND Characterization and Experiment Setup

Electrical measurements were performed on five TLC and three different quad-level-cell (QLC) 3-D NAND Flash products featuring over 150 stacked layers (N > 150). All program, erase, and read operations were conducted in TLC or QLC mode. Device testing encompassed all page types—including the least significant bit (LSB), central significant bit (CSB), and most significant bit (MSB) for TLC Flash, as well

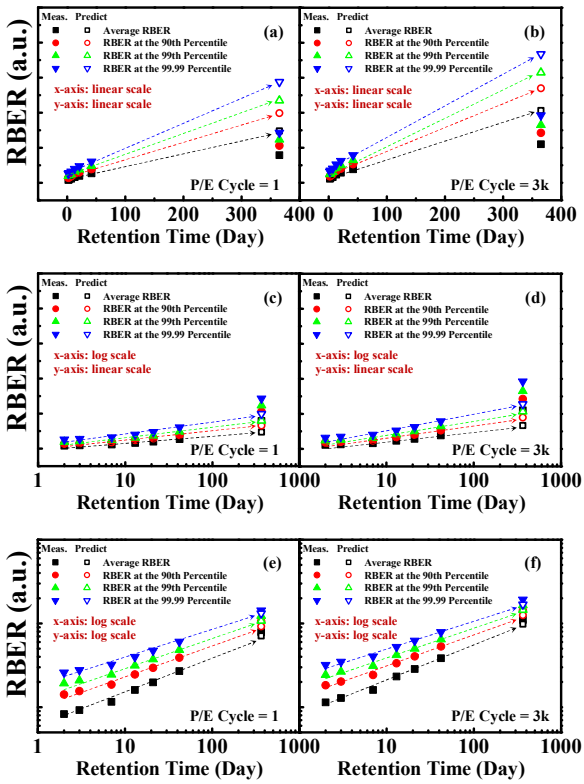

Fig. 4. Evolutions of RBER with retention time at P/E cycle = 1 and 3k exhibited in linear-linear scale, log-linear scale, and log-log scale. Flash type is TLC NAND from vendor A.

as the LSB, central least significant bit (CLSB), central most significant bit (CMSB), and MSB for QLC Flash—across all word-line layers. The RBER was evaluated by applying 7 and 15 optimized read voltages for the TLC and QLC Flash memories, respectively.

Before retention characterization, the blocks were first P/E cycled for 1 and 3k cycles using an incremental step pulse P/E scheme with random data patterns at a temperature of T = 85°C. Subsequently, the devices were erased again, followed by a program operation with random data patterns at room temperature, and then stored for retention at 30°C or 40°C. The RBER was continuously monitored using an SSD-controller-based raw NAND test platform through controller-issued read operations ranging from a few days to 1 year, as shown in Fig. 3. To reduce manufacturing-induced variability, measurements were performed across multiple memory blocks and dies within the same product. The disturbance induced by the applied voltages during measurements was carefully considered to avoid its impact on the results.

III. Results and Discussion

A. Comparison between the 1-year retention data from actual measurement result and from the three prediction models

To evaluate which prediction method provides higher precision, we first selected a type of TLC NAND device from vendor A for analysis. The one-year retention data obtained from empirical measurements was compared against the predictions of three distinct models to evaluate their accuracy.

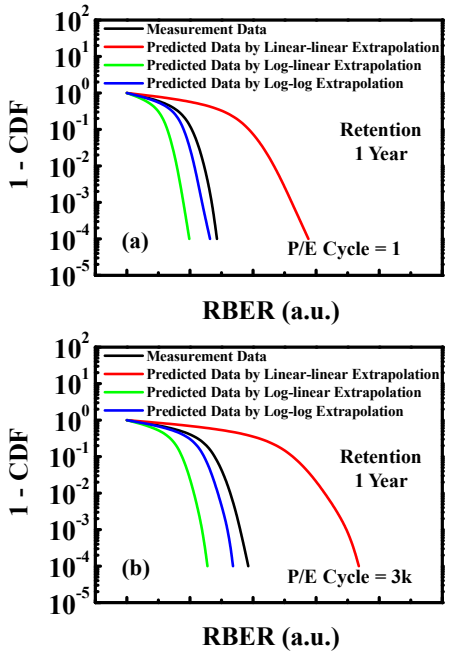

Fig. 5. Measurement and 3 different predicted 1-CDF of RBER. Retention time is 1 year. Flash type is TLC NAND from vendor A.

Fig. 4 presents the evolution of the average RBER, as well as the RBER at the 90th, 99th, and 99.99th percentiles, as a function of retention time under P/E cycle counts of 1 and 3k. These subfigures employ different combinations of linear and logarithmic axes (specifically, Fig. 4(a) and Fig. 4(b) use a linear–linear scale; Fig. 4(c) and Fig. 4(d) use a log–linear scale; and Fig. 4(e) and Fig. 4(f) use a log–log scale) to evaluate both the fitted trend lines and the one-year retention extrapolation. In the plots, the dashed lines represent the fitted straight line, while the hollow markers indicate the RBER values extrapolated to one year of retention. As shown in Fig. 4, the log–log extrapolation model representations in Fig. 4(e) and Fig. 4(f) provide the most accurate prediction results. For the other two prediction approaches, the linear–linear extrapolation overestimates the growth of RBER, whereas the log–linear extrapolation underestimates it. Fig. 5 presents the one-year retention and the corresponding 1-CDF of RBER predicted by the three models, where the trends remain consistent with those observed in Fig. 4. Among the three models, the one-year retention predicted by the log–log extrapolation shows the closest agreement with the actual data based on our measurement results. This conclusion is consistently observed across various other NAND Flash types (not shown here), further supporting its general applicability.

B. Physical insight and justification of the log–log extrapolation

The following discussion outlines the rationale behind using log-log extrapolation to provide the most accurate predictive performance. The data written into each memory cell were organized into 8 and 16 different V_t states for TLC and QLC Flash, respectively. The overlap between two neighboring states and their corresponding parameters at the beginning of retention (t=0) and at a specific retention time t are defined as

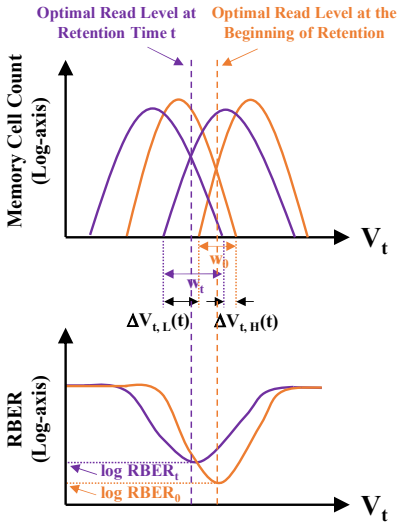

Fig. 6. Schematic Diagram of V_t and RBER distribution at the beginning of retention (orange curve) and at retention time t (purple curve). w_0, w_t, $\Delta V_{t,L}(t)$ and $\Delta V_{t,H}(t)$, log $RBER_t$, and log $RBER_0$ are defined.

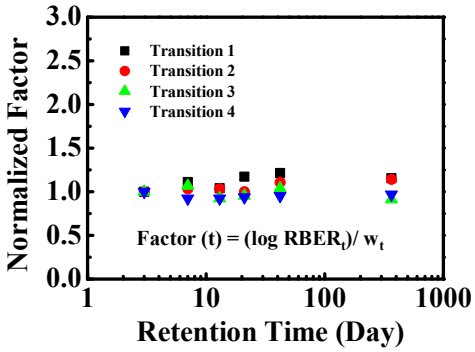

Fig. 7. Evolutions of normalized factor with retention time for several transitions across different NAND types. Normalized factor is defined as Factor (t) / Factor (0).

illustrated in Fig. 6. $RBER_t$ was obtained by adjusting the read voltage to achieve the minimum bit error, namely the optimal read voltage, at retention time t. Following the analysis of over ten thousand transitions across various NAND Flash types, we observed that the factor $(\log RBER_t)/w_t$ remained almost unchanged during retention, as shown in Fig. 7. Since the factor defined in Fig. 7 does not vary with retention, the equation can be expressed as follows,

$$\frac{\log RBER_0}{w_0} = \frac{\log RBER_t}{w_t} \quad (1)$$

where the parameters are defined in Fig. 6, and w_0 and w_t satisfy the following relationship:

$$w_t = w_0 + \Delta V_{t,L}(t) - \Delta V_{t,H}(t). \quad (2)$$

Furthermore, according to our previous study [9], room-temperature retention loss in NAND Flash memory is governed by trapped-electron vertical loss via direct tunneling, irrespective of whether the devices are fresh cells or P/E-cycled. This physical phenomenon is typically described by the tunneling-front model [18], and the resulting V_t shift exhibits a logarithmic time dependence, as given by:

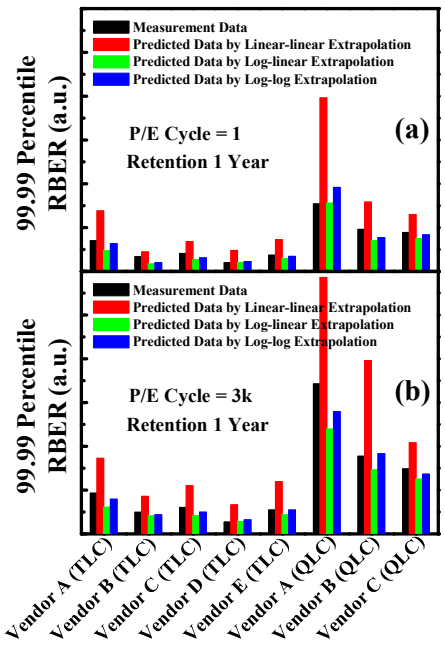

Fig. 8. Measurement and 3 different predicted 99.99 percentile RBER with 8 different types (5 TLC and 3 QLC) of NAND Flash. Retention time is 1 year.

$$\Delta V_{t,L}(\mathrm{t}) \propto \log t \tag{3}$$

and

$$\Delta V_{t,H}(\mathrm{t}) \propto \log t\ . \tag{4}$$

Therefore, from eq.(1) to eq.(4), the following equation can be derived, as

$$\log RBER_t \propto \log t\ . \tag{5}$$

Based on the aforementioned physical insights and justification, the log–log extrapolation approach given in Eq. (5) demonstrates the highest efficacy as a predictive model. Crucially, this capability extends beyond the TLC NAND Flash from Vendor A. By summarizing the retention results across all eight distinct NAND Flash devices, the log–log extrapolation framework is found to consistently yield predictions that closely align with the empirical retention data, as demonstrated in Fig. 8.

IV. Conclusions

While prior studies lacked sufficient data to evaluate predictive models, this work established a long-term, one-year retention database across multiple NAND Flash types. The measurement results confirm that log–log extrapolation delivers the most accurate prediction for room-temperature retention behavior. Moreover, the physical basis for the superior accuracy of this log–log approach is comprehensively explained.

References

[1] H. Tanaka *et al.*, "Bit cost scalable technology with punch and plug process for ultra high density Flash memory," in *Proc. IEEE Symp. VLSI Technol.*, Jun. 2007, pp. 14−15, doi: 10.1109/VLSIT.2007.4339708.

[2] J. Jang *et al.*, "Vertical cell array using TCAT(terabit cell array transistor) technology for ultra high density NAND Flash memory," in *Proc. IEEE Symp. VLSI Technol.*, Jun. 2009, pp. 192−193.

[3] R. Micheloni, S. Aritome, and L. Crippa, "Array architectures for 3-D NAND Flash memories," *Proc. IEEE*, vol. 105, no. 9, pp. 1634–1649, Sep. 2017, doi: 10.1109/JPROC.2017.2697000.

[4] H.-J. Kang, *et al.*, "Comprehensive analysis of retention characteristics in 3-D NAND flash memory cells with tube-type poly- Si channel structure," in *Proc. Symp. VLSI Technol. (VLSI Technol.)*, Jun. 2015, pp. T182–T183, doi: 10.1109/VLSIT.2015.7223670.

[5] B. Choi, *et al.*, "Comprehensive evaluation of early retention (fast charge loss within a few seconds) characteristics in tube-type 3-D NAND Flash memory," in *Proc. IEEE Symp. VLSI Technol.*, Jun. 2016, pp. 78−79, doi: 10.1109/VLSIT.2016.7573385.

[6] W. J. Tsai, *et al.*, "Data retention behavior of a SONOS type two-bit storage flash memory cell," in *IEDM Tech. Dig.*, Dec. 2001, pp. 32.6.1–32.6.4, doi: 10.1109/IEDM.2001. 979614.

[7] L. Breuil, *et al.*, "Comparative reliability investigation of different nitride based local charge trapping memory devices," in *Proc. IEEE Int. Rel. Phy. Symp. (IRPS)*, Apr. 2005, pp. 181–185, doi: 10.1109/RELPHY.2005.1493081.

[8] H. T. Lue, *et al.*, "Study of charge loss mechanism of SONOS-type devices using hot-hole erase and methods to improve the charge retention," in *Proc. IEEE Int. Rel. Phy. Symp. (IRPS)*, Mar. 2006, pp. 523–529. doi: 10.1109/RELPHY.2006.251273.

[9] Y. H. Liu, T. C. Zhan, Y. S. Yang, C. C. Hsu, A. C. Liu, and W. Lin, "Impact of trapped charge vertical loss and lateral migration on lifetime estimation of 3-D NAND Flash memories," in *Proc. IEEE Int. Rel. Phy. Symp. (IRPS)*, Mar. 2023, pp. 10C.4-1–10C.4-6. doi: 10.1109/IRPS48203.2023.10118289.

[10] D. A. Baglee, "Characteristics & reliability of 100 Å oxides," in *Proc. IEEE Int. Rel. Phy. Symp. (IRPS)*, Apr. 1984 pp. 152–155. doi: 10.1109/IRPS.1984.362035.

[11] K. Lee, M. Kang, S. Seo, D. H. Li, J. Kim, and H. Shin, "Analysis of failure mechanisms and extraction of activation energies (E_a) in 21-nm NAND Flash cells," *IEEE Electron Device Lett.*, vol. 34, no. 1, pp. 48−50, Jan. 2013, doi: 10.1109/LED.2012.2222013.

[12] P. Kumari, U. Surendranathan, M. Wasiolek, K. Hattar, N. Bhat, and B. Ray, "Analytical bit-error model of NAND Flash memories for dosimetry application," *IEEE Trans. Nuclear Science*, vol. 69, no. 3, pp. 478−484, Mar. 2022, doi: 10.1109/TNS.2021.3125652.

[13] Y. Cai, Y. Luo, E. F. Haratsch, K. Mai, and O. Mutlu, "Data retention in MLC NAND flash memory: Characterization, optimization, and recovery," in *Proc. IEEE 21st Int. Symp. High-Perform. Comput. Archit. (HPCA)*, Feb. 2015, pp. 551–563, doi: 10.1109/HPCA.2015.7056062.

[14] D. Wei, L. Qiao, X. Chen, H. Feng, and X. Peng, "Prediction models of bit errors for NAND flash memory using 200 days of measured data," *Rev. Sci. Instrum.*, vol. 90, no. 6, Jun. 2019, Art. no. 064702, doi: 10.1063/1.5064655.

[15] Y. Kim and S. Kim, "Simulation acceleration of bit error rate prediction and yield optimization of 3D V-NAND Flash memory," *IEEE Access*, vol. 11, pp. 93956–93967, 2023, doi: 10.1109/ACCESS.2023.3309649.

[16] J. He, Q. Zhao, X. Qiang, Q. Wang, Q. Li, and Z. Huo, "Hybrid neural network model for raw bit error rate prediction of 3-D TLC NAND Flash memory," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 44, no. 12, pp. 4657–4666, Dec. 2025, doi: 10.1109/TCAD.2025.3569487.

[17] H. Tian *et al.*, "Modeling retention errors of 3D NAND Flash for optimizing data placement," *ACM Trans. Design Autom. Electron. Syst.*, vol. 29, no. 4, Jun. 2024, Art. no. 62, doi: 10.1145/3659101.

[18] S. Manzini and A. Modelli, "Tunneling discharge of trapped holes in silicon dioxide", in insulation films on semiconductors, J. F. Verwij and Wolters, Eds. Amsterdam, The Netherlands; Elsevier, p. 112, 1983.