Endurance Characteristics of SuperFlash® Memory

Xian Liu*, Viktor Markov, Alexander Kotov, Tho Ngoc Dang, Amitay Levi
Silicon Storage Technology, Inc., 1171 Sonora Court, Sunnyvale, CA94086, USA
Ian Yue, Andy Wang, and Rodger Qian
SST China, Ltd., Bldg. 24, No.115, Lane 572, Bibo Road, Zhangjiang Hi-Tech Park, Shanghai, 201203, P.R. China
* Email: xiliu@sst.com

Abstract

Program/erase endurance characteristics of split-gate SuperFlash® memory cells are discussed. Various factors which affect memory endurance, including cycling data pattern, cycling frequency, temperature, erase retries, and technology scaling, are investigated. Superior data retention after endurance cycling is demonstrated.

1. Introduction

The SuperFlash® (Figure 1) is a split-gate flash memory cell utilizing source-side channel hot electron injection for program, and poly-to-poly Fowler-Nordheim (F-N) tunneling for erase [1],[2].

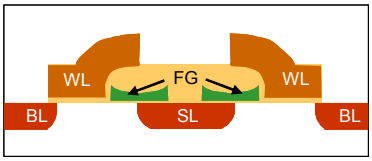

Figure 1a. Non-self-aligned SuperFlash® cell for low-density flash memory.

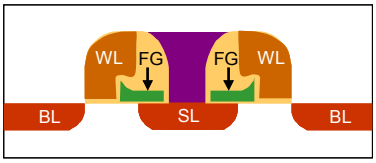

Figure 1b. Self-aligned SuperFlash® cell for medium- and high-density flash memory.

During erase, a high voltage is applied on the word-line (WL) to induce F-N tunneling of electrons from the floating gate (FG) to the WL. Instead of exploiting tunneling through a flat thin oxide layer [3], as commonly adopted by stacked-gate cells, the SuperFlash® utilizes local tunneling field enhancement near a tip formed on the FG, which promotes the tunneling of electrons. This geometrical field enhancement enables the use of relatively thick tunnel oxides operating under relatively low average fields. Such tunnel oxides are more resistant to stress-induced leakage current (SILC) [4], and are easier to scale.

With the suppression of SILC, electron trapping, particularly in the tunnel oxide, becomes the main cause of reliability failures. Charge trapping during erase occurs both through the filling of existing traps and through the generation and filling of new traps by tunneling electrons. The trapped electrons reduce the F-N tunneling field and thus the efficiency of subsequent erases. Such erase weakness is the dominant endurance failure mode observed in SuperFlash® cells. The failures are not catastrophic, but can be alleviated by using stronger erase, by optimizing operating conditions to reduce charge trapping, and through the relaxation of trapped charges during quiescent periods.

It should be pointed out that, during program, some electron trapping does happen in the oxide beneath the FG (FGOX), which also leads to reduction of "1" state read current. Such mechanism has been reported to cause endurance failures [5]. However, our studies, the details of which are beyond the scope of the present paper, indicate that trapping in the FGOX exerts much smaller effect on endurance than trapping in the tunnel oxide.

The following sections discuss endurance characteristics of SuperFlash® memory. Effects of cycling conditions which affect endurance, including the data pattern used in cycling tests, cycling frequency, and temperature are reported. Effects of erase retries and technology scaling are also included. Finally, data retention after endurance cycling is discussed.

2. Endurance Characteristics

2.1 Data pattern

As endurance is limited by electron trapping in the tunnel oxide, it is directly affected by the amount of charge transferred through this oxide during each erase operation. If a cell is written with data "0" during program, then during the subsequent erase, a certain amount of electrons has to be tunneled in order to bring the cell to the "1" state. Some of these electrons can be trapped in the oxide. On the other hand, if the cell is written with data "1" during program, then tunneling current during the subsequent erase is much lower, and trapping is relatively insignificant. Figure 2 shows a comparison of endurance lifetime distributions of two groups of cells from the same memory array, each cycled with a different data pattern. The use of a rotating mixed data pattern, which more closely resembles the random data used in typical application conditions, gives rise to 2X improvement in endurance lifetime as compared to the case where each cell is programmed to "0" state in

each cycle.

2.2 Relaxation effects

Endurance testing is typically performed with continuous program and erase operations to reduce test time. This is in contrast to the infrequent cycling occurring in typical usage. In the latter case, relaxation of trapped charges occurs during the quiescent periods between erase and program operations, leading to additional endurance improvements. It has been estimated that the endurance of SuperFlash® in typical applications (average cycling frequency < 1 cycle/hr) is at least 2X greater than the endurance demonstrated in a stress environment using maximum frequency [6].

Relaxation of trapped charges is accelerated at elevated temperatures. A comparison of endurance distributions under different cycling temperatures in [6] shows more than 2X endurance improvement at 125 °C as compared to room temperature result. High-temperature endurance is particularly important in automotive applications.

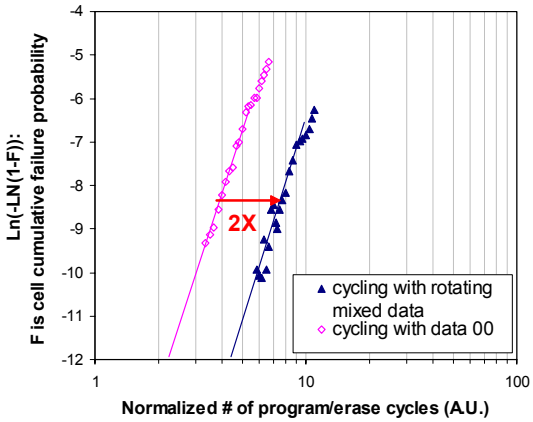

Figure 2. Weibull distributions of endurance lifetime from two 32Kb sectors P/E cycled with different data patterns: *rotating mixed data* (each program cycle writes 50% of cells to "0" and leave the rest at "1". The data pattern is inverted in adjacent write cycles, so that each cell is programmed to "0" state in half of the write cycles) and *Data 00* (all cells are programmed to "0" in every cycle). Failure rates in both groups are based on "1"-state read failures after 1-time erase from "0" state.

2.3 Erase retries

Endurance failures caused by electron trapping in the tunnel oxide are not catastrophic, but can be alleviated by increasing the strength of erase. An efficient way to improve erase strength and reduce endurance failures is the use of erase retries. In this scheme, one can use memory cell verification after each erase operation, and apply more erase pulses as needed. Figure 3 shows endurance lifetime distributions of a memory array with the use of erase retries. It is observed that every doubling of erase time gives rise to roughly 1.5-1.7X improvement in endurance lifetime. Memory arrays undergone extensive endurance testing with up to 10 erase retries did not exhibit any new failure mode related to extended erase stress.

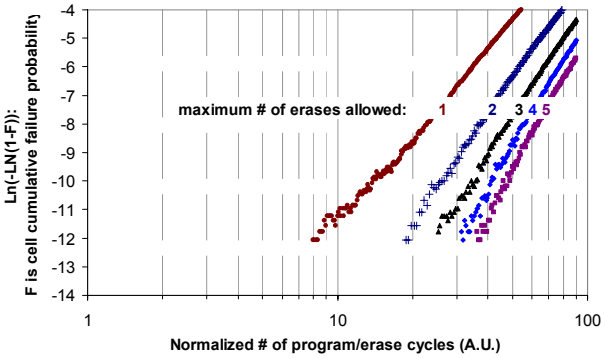

Figure 3. Endurance lifetime distributions of a memory array with the use of erase retries. During cycling, cells are verified after each erase, and erase is repeated as needed until all cells in the sector pass verification.

2.4 Technology scaling

It has been previously demonstrated that tunneling degradation due to electron trapping in the tunnel oxide is reduced as the tunnel oxide thickness is scaled [6]. This effect can be in part explained by the electrical screening effect. The highest trapped-electron density is expected to occur in a thin slice of the tunnel oxide located just beyond the tunneling barrier [7]. These charges are also the most effective in screening the tunneling field during erase. As the oxide becomes thinner, the effectiveness of these trapped electrons in screening the tunneling field is reduced. Furthermore, thinner oxides are also expected to trap fewer electrons, as a result of the reduced distance of interaction between the tunneling electron and the oxide. Both of these effects explain the reduction of trapping-related erase degradation in thinner tunneling oxides.

Technology scaling benefits endurance not only through oxide scaling, but through scaling of lateral dimensions as well. Scaling the cell in both length and width directions leads to faster scaling of FG capacitance than the scaling of tunneling edge length. This in turn leads to a reduction of tunneling current density, and hence a reduction of electron trapping. Figure 4 shows endurance lifetime distributions from three generations of the SuperFlash®. The tunnel oxide thickness has been scaled together with lateral dimensions. A clear correlation between technology scaling and endurance improvement is evident. The typical endurance lifetime of a 16Mb memory using 0.18um technology is estimated to be 1

million cycles under typical application conditions. Further endurance improvement can be achieved by using error correction codes (ECC).

3. Retention after endurance cycling

In general, the endurance of floating-gate memories is not only limited by reduction of program/erase efficiency, but also by data retention failures related to SILC [4]. In SuperFlash® cells, the average field in the tunnel oxide during erase is about 6-7MV/cm. This is significantly lower than the 10MV/cm used in stacked-gate cells. In addition, the high erase efficiency translates to significantly shorter erase time. The thinner oxide underneath the FG is not exposed to high field, hence is more resistant to SILC than its counterpart in stacked-gate cells. These advantages make the SuperFlash® more resistant to cycling-related retention failures. Figure 5 shows cell read current distributions in a 32Kb array which was cycled extensively (3million cycles) and subsequently baked at 150 °C for 1500 hours. A comparison of the distribution before and after baking indicates the absence of leaky bits even after such excessive cycling. (After baking, the distribution is slightly and uniformly shifted toward higher current. This has been attributed to relaxation of electrons trapped in the FGOX.) Similar experiments using lower-temperature bakes for longer time durations show the same trend.

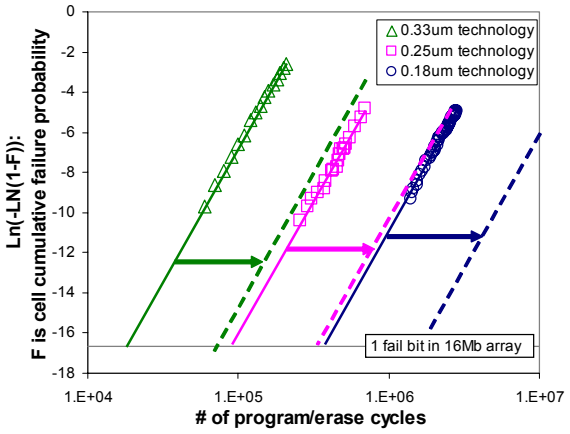

Figure 4. Endurance lifetime distributions from 3 SuperFlash® generations. Data points (solid lines) were measured at room temperature using continuous P/E cycling of all cells to accelerate testing. Erase retry was not used. Dashed lines corresponds to the minimum 4X improvement expected under typical application conditions, taking into account differences in cycling data pattern and frequency.

4. Summary

The program/erase endurance characteristics of split-gate SuperFlash® memory cells are strongly affected by cycling conditions, including the data pattern, cycling frequency, and temperature. These effects have been quantified and can be combined with cycling stress test results to estimate endurance lifetime under specific application conditions. Erase retries have been demonstrated to effectively improve endurance lifetime. Continuous endurance improvement has been achieved through technology scaling. A typical endurance lifetime of 1 million cycles is expected in the 0.18um technology. Superior data retention after endurance cycling is demonstrated.

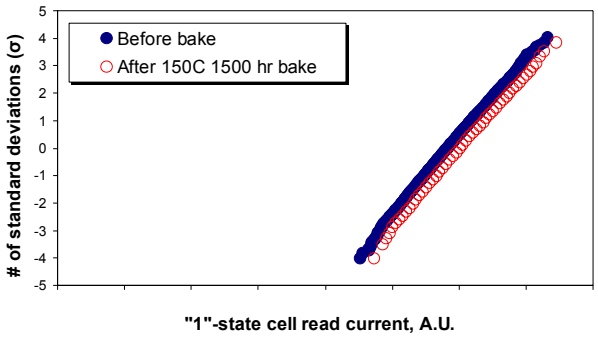

Figure 5. Normal probability plot for "1"-state cell read current of 32Kb array subjected to 3 million P/E cycles. Distribution was measured before and after 1500 hours bake at 150 °C.

Acknowledgments

The authors would like to thank Yuri Tkachev for many insightful discussions, Austin Huang for data collection, and Liz Cuevas for proof-reading the manuscript.

References:

[1] S. Kianian, A. Levi, D. Lee, and Y.-W. Hu, Proc. Symp. VLSI Technol., p.71 (1994).

[2] R. Mih, J. Harrington, K. Houlihan, H.K. Lee, K. Chan, J. Johnson, B. Chen, J. Yan, A. Schmidt, C. Gruensfelder, K. Kim, D. Shum, C. Lo, D. Lee, A. Levi, C. Lam, Symp. VLSI, p.120 (2000).

[3] M. Lenzlinger and E.H. Snow, J. Appl. Phys., 40, p.278 (1969).

[4] J. de Blauwe, J. Van Houdt, D. Wellekens, G. Groeseneken, H.E. Maes, IEEE Trans. Electron Dev., 45, p.1745 (1998).

[5] T.I. Wu, Y.D. Chih, S.H. Chen, W. Wang, Mi-Chang Chang, J.R. Shih, H.W. Chin, K. Wu, IPFA, p. 115 (2004).

[6] V. Markov, X. Liu, A. Kotov, A. Levi, T.N. Dang, and Y. Tkachev, NVMTS, p.23-1 (2003).

[7] A. Kotov, Mat. Res. Soc. Symp. Proc. 473, p.185 (1997).