A New Combination–Erase Technique for Erasing Nitride Based (SONOS) Nonvolatile Memories

Gowrishankar L. Chindalore, *Member, IEEE*, Craig T. Swift, and David Burnett, *Member, IEEE*

***Abstract*—A new technique of erasing nonvolatile memory (NVM) devices based on nitride storage (SONOS) with bottom oxide thickness in the range of 30 Å has been developed. Oxide thickness in this range is necessary to minimize the undesirable effects of gate disturb while still enabling a low-voltage operation to maximize the cost benefit of SONOS memories. To erase such bitcells, Fowler–Nordheim tunneling (FNT) is preferred over hot-hole injection (HHI) due to the less damaging nature of FNT. However, FNT alone cannot be used to erase the device completely due to erase saturation limitations. Hence, the new "combination–erase" technique combines both FNT and HHI erase to achieve a fast and controlled erase. Furthermore, by using FNT erase at higher field conditions, and HHI erase at lower field conditions, the reliability of the bitcell is also improved.**

***Index Terms*—Combination–erase, data retention, Fowler–Nordheim tunneling (FNT), gate disturb, hot-electron injection, hot-hole injection (HHI), nitride traps, SONOS memory.**

I. Introduction

NONVOLATILE memories (NVM) based on storage in nitride traps have gained increased interest due to the simplicity of the bitcell structure and process, low-voltage operation, and its immunity to extrinsic charge loss as compared to traditional floating gate (FG)-based NVM. Nitride-based memories can be further broadly classified into nonlocal charge-storage memories, such as SONOS, and localized charge-storage memories, such as those popularized by NROM [1]. In either case, the nonvolatility is achieved by storing charge in nitride traps and surrounding the nitride with oxide to form an oxide–nitride–oxide (ONO) stack. The more well-known SONOS-based memories with thin bottom oxide (<25 Å) typically utilize uniform tunneling for both program and erase operations resulting in programming times in the 1-ms range, which is slower than desired for many high-density, embedded Flash electrically erasable programmable read-only memory (EEPROM) applications [2], [3]. Furthermore, the use of thin oxides causes severe read disturb in the selected bitcell and gate disturb in the unselected bitcells sharing the same word line during a read operation. An architectural approach to minimize the gate disturb for SONOS memories has been developed by Lancaster *et al.* [4], but the programming speed limitation is still the major factor that has prevented widespread use of these thin-oxide SONOS devices.

Manuscript received January 2, 2003. The review of this letter was arranged by Editor S. Chung.

The authors are with the Embedded Memory Center, Technology and Manufacturing, Semiconductor Products Sector, Motorola, Inc. Austin, TX 78735 USA.

Digital Object Identifier 10.1109/LED.2003.810883

On the other hand, localized charge-storage memories are typically programmed using channel hot-electron injection (HEI) and erased using hot holes generated by band-to-band tunneling in the source and drain overlap regions. In this case, both the program and erase are very fast operations, but require the use of much thicker bottom oxide in the order of 70 Å [5] in order to minimize damage to the bottom oxide by energetic holes during erase operation. Even with such thick oxides, read and gate disturb from the low V_t state, and degradation in the data retention characteristics from the high V_t state, have been shown to occur after thousands of program/erase (P/E) cycles [5], [6]. Additionally, the use of thick oxides requires high voltages (greater than 9 V) for program and erase operations. To support these high voltages, large peripheral transistors must be used, thus decreasing the cost benefit offered by the use of SONOS memories.

From the above discussion, it is clear that SONOS-based memories using FNT erase are preferable if the programming speed of such bitcells can be increased and gate disturb minimized. To that end, we recently reported a SONOS memory device with thin bottom oxide (22 Å) utilizing HEI for programming and FNT for erase [7], with programming speeds in the range of 1–10 μs, and erase speeds in the range of 100 ms. The uncycled gate disturb of the bitcell has also been shown to substantially decrease by lowering the natural threshold voltage $(V_{t,\mathrm{nat}})$ of the device, but the cycled gate disturb remains a concern. By decreasing the $(V_{t,\mathrm{nat}})$ of the device, the vertical field during gate disturb is decreased as governed by the equation for the bottom oxide field $E = (V_g + V_{t,\mathrm{nat}} - V_{t,\mathrm{erase}})/T_{\mathrm{eff}}$, where T_{eff} is the effective ONO thickness. One model that explains the gate disturb enhancement after P/E cycling is that the cycling creates electron and interface traps that enhance hole tunneling from nitride into the channel. Thus, an easy way to improve post-cycled gate disturb would be to increase the tunneling distance between the holes and the oxide traps by increasing the bottom oxide thickness to approximately 30 Å. However, with increasing bottom oxide thickness, the erase speed also decreases exponentially, as governed by electric field dependence. In this thickness range, the FNT erase itself is severely limited by the "erase saturation" effect, which occurs when the electrons are injected from the gate into the nitride faster than they can be removed through the bottom oxide due to increasing top oxide field.

II. New Technique

In order to overcome the problem of erase saturation associated with erasing bitcells with intermediate oxide thickness

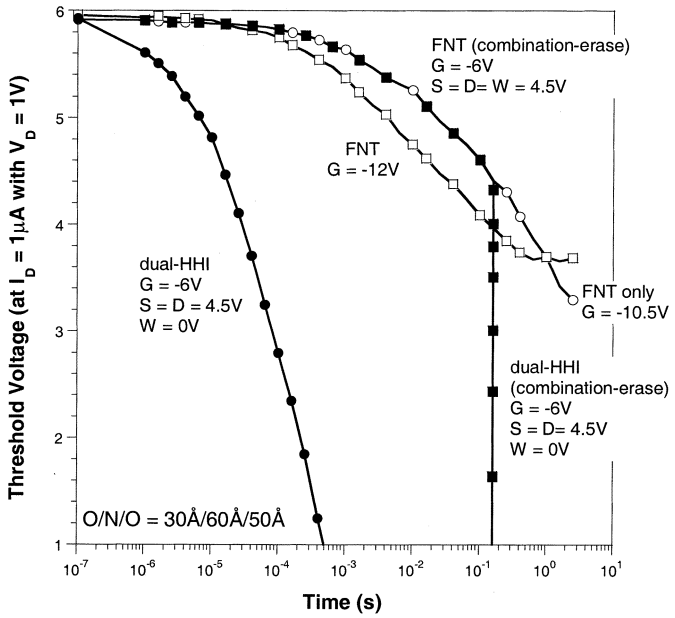

Fig. 1. Erase speed comparison between FNT, HHI, and combination–erase techniques. FNT erase saturates quickly with higher gate biases. The total erase time for combination–erase is approximately the same as FNT erase time, since HHI is substantially faster than FNT erase.

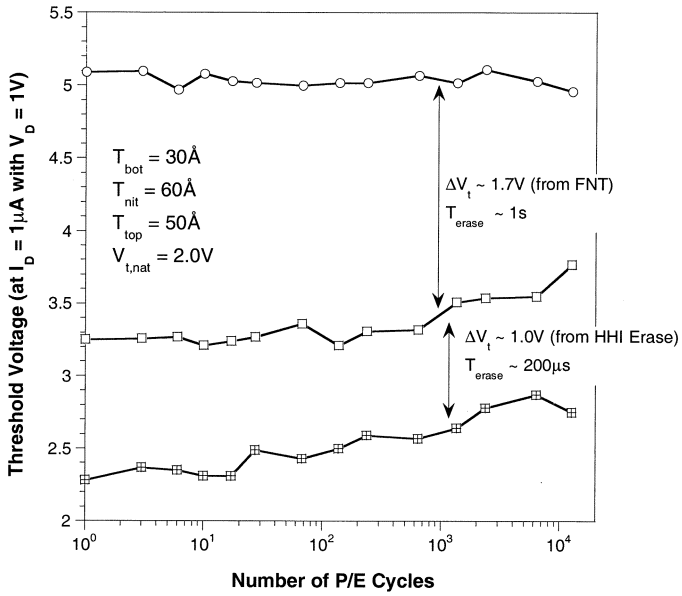

Fig. 2. Endurance characteristics of SONOS bitcells with 30 Å bottom oxide thickness that is erased using combination–erase. The erase V_t increase is due to transconductance degradation caused by oxide and interface traps formed during HCI programming.

(30 Å), we propose a new two-step erase technique called "combination–erase." In this technique, the bitcell is initially erased using FNT to the onset of saturation, and then the remainder of the erase is completed using source and drain HHI. During FNT, the total erase voltage is split between the gate and the well/source/drain in order to enable low-voltage erase operation. During the subsequent HHI erase, the well bias is turned off while maintaining a negative gate bias, and positive source and drain biases. As a result, electron-hole pairs created from band-to-band tunneling in the source/drain overlap regions are accelerated by the lateral electric field and injected into the nitride assisted by the vertical electric field to complete the erase. It should be noted that the ONO stack has to be optimized in order to maximize the FNT erase by suppressing saturation and minimize the extent to which HHI erase is used. In this letter, we report the implementation of this technique in SONOS-based memories for 1 bit/cell storage, but the technique can also be equally well applied to localized charge storage memories to achieve 2 bits/cell.

III. Results and Discussion

SONOS memory devices with bottom oxide thicknesses in the range of 22–30 Å, low-pressure chemical vapor deposition (LPCVD) nitride with thicknesses ranging from 70–150 Å, and top oxide thicknesses of approximately 50 Å that are either deposited or thermally grown have been processed with channel lengths ranging from 0.16 to 0.26 μm using 90-nm advanced copper technology [7].

The erase speed of FNT, HHI erase, and the combination–erase are compared in Fig. 1. The FNT erase with -12 V gate bias is fast at the beginning of erase, but the erase V_t quickly saturates at approximately 4 V. The HHI erase is undoubtedly faster and more efficient, but imparts damage to the bottom oxide, as will be shown later. In the combination–erase case, the total erase time is approximately the same as the FNT erase time, since the HHI erase time is negligibly small compared to FNT erase time. Fig. 2 shows the endurance characteristics of combination–erase with voltage conditions same as Fig. 1. The increase in erase threshold voltage observed during cycling is due to transconductance and subthreshold slope degradation caused by the oxide and interface traps generated by hot electrons during programming [8], [9]. Similarly, we observe an increasing degradation of the transconductance and subthreshold slope with cycling that correlates with the increase in erase V_t. The increasing erase V_t is not an indication of an electron build-up, since the V_t also increases after the FNT portion of the combination erase, but a uniform tunnel erase in itself is resistant to electron build-up in the nitride. If the electrons were to build up in the nitride, then the local vertical field also increases correspondingly, resulting in tunneling of those excess electrons.

Fig. 3 compares data retention at 150 °C between bitcells that are cycled to 10 000 times using HCI/FNT, HCI/combination–erase, and HCI/HHI. The bottom oxide thickness is 22 Å for the HCI/FNT cycling and 30 Å for HCI/combination–erase and HCI/HHI cycling. Furthermore, we have also experimentally determined that the difference in the nitride thickness between the samples has no effect on the data retention (DR) behavior. The DR comparison shows that both combination–erase and HHI erase are worse than FNT erase, but the combination–erase is substantially better than HHI erase over a longer period of time. This inherent improvement in the reliability with combination–erase is achieved by using FNT erase when the oxide field is the highest and using HHI erase when the oxide field is much lower.

The gate disturb improvement achieved by using 30-Å bottom oxide thickness is shown in Fig. 4. This figure shows

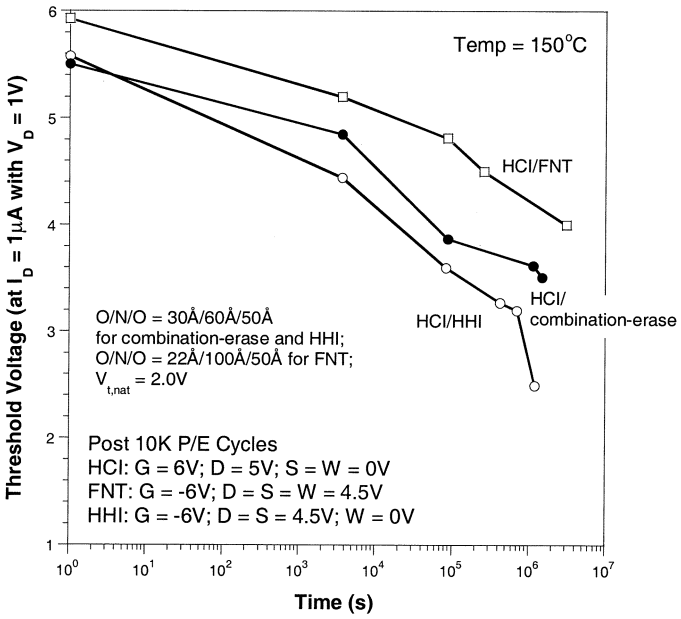

Fig. 3. High-temperature data retention comparison between bitcells that are P/E cycled 10 K times using various erase mechanisms. The combination–erase is found to be more reliable than HHI erase alone over a longer period of time.

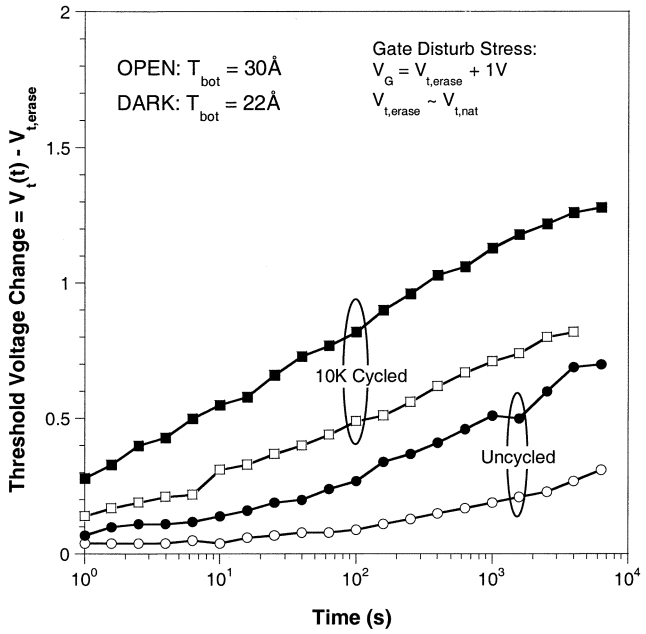

Fig. 4. Gate disturb dependence on oxide thickness is strongly evident, especially after P/E cycling for bitcells with a high natural threshold ($V_{t,\mathrm{nat}} \sim 2$ V). Combination–erase enables use of 30-Å bottom oxide that has much less gate disturb than thin oxides.

substantial degradation of gate disturb after cycling in bitcells with 22 Å bottom oxide as compared to bitcells with 30 Å bottom oxide. The gate bias for this disturb measurement is chosen to be 1 V higher than the erase V_t in order to ensure sufficient read current, and in order to enable a fast read access ($<$15 ns). The gate disturb degradation after cycling is believed to be caused by the electron and interface traps that are formed during cycling, which not only tend to increase the natural threshold of the device, but also enhance hole emission from the nitride [10], [11]. With a thicker bottom oxide, the tunneling distance is increased, and the hole emission probability is considerably decreased. Furthermore, we have reported that the gate disturb can be dramatically decreased by decreasing the natural threshold voltage of the bitcell [7]. Thus, combining these two results, it is possible to design an optimized SONOS bitcell with a low natural threshold and bottom oxide thickness in the range of 30 Å that can be programmed and erased with low voltages, has good reliability, and suffers from little or no gate disturb after cycling. Another circuit-related advantage of combination–erase over HHI erase is that the erase current requirement of combination–erase is lower than HHI erase. As a result, the charge pumps supplying the erase voltages can be smaller to further enhance the cost benefit of SONOS memories.

IV. Conclusions

A new combination–erase erase technique that combines the benefits of FNT erase and the HHI erase has been demonstrated for use in SONOS NVM devices with an intermediate range of bottom oxide thickness (>30 Å). Very thin bottom oxides are limited by gate disturb after thousands of cycling, and very thick bottom oxides require high voltages and are subject to hot-hole induced damage to the bottom oxide. However, the combination–erase technique has been shown to enable development of SONOS bitcells that have fast program and erase speeds, are relatively disturb-free, and can be operated with low voltages. In addition, the reliability of the SONOS bitcells is also improved by utilizing FNT erase when the erase fields are high and by using HHI for only a short period of time under lower erase field conditions.

References

[1] B. Eitan, P. Pavan, I. Bloom, E. Aloni, A. Fromeer, and D. Finzi, "NROM: A novel localized trapping, 2-bit nonvolatile memory cell," *IEEE Electron Device Lett.*, vol. 21, pp. 543–545, Nov. 2000.

[2] S. Minami and Y. Kamigaki, "A novel monos nonvolatile memory device ensuring 10-year data retention after 10^7 erase/write cycles," *IEEE Trans. Electron Devices*, vol. 40, pp. 2011–2017, Nov. 1993.

[3] A. Nakamura, H. Aozasa, I. Fujiwara, M. Yamagishi, and Y. Komatsu, "The characteristics of sub-quarter-μm MONOS memory transistor with a thicker tunnel oxide as a single transistor cell," in *Proc. IEEE Non-Volatile Silicon Memory Workshop*, Feb. 1997, Dig. 3.4.

[4] L. Lancaster, "Why SONOS?," in *Proc. IEEE Non-Volatile Silicon Memory Workshop*, Aug. 2001, pp. 24–26.

[5] 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.

[6] W. J. Tsai *et al.*, "Hot carrier enhanced read disturb and scaling effects in a localized trapping storage SONOS type flash memory cell," in *Proc. Int. Conf. Solid State Devices Mat.*, Nagoya, Japan, 2002, pp. 164–165.

[7] C. T. Swift *et al.*, "An embedded 90 nm SONOS nonvolatile memory utilizing hot electron programming and uniform tunnel erase," in *IEDM Tech. Dig.*, Dec. 2002, pp. 927–930.

[8] S. Yamada *et al.*, "Degradation mechanism of flash EEPROM programming after program/erase cycles," in *IEDM Tech. Dig.*, Dec. 1993, pp. 23–26.

[9] S. S. Chung, C.-M. Yih, S.-M. Cheng, and M.-S. Liang, "A new technique for hot carrier reliability evaluations of Flash memory cell after long-term program/erase cycles," *IEEE Trans. Electron Devices*, vol. 46, pp. 1883–1889, Sept. 1999.

[10] K. I. Lundstrom and C. M. Svensson, "Properties of MNOS structures," *IEEE Trans. Electron Devices*, vol. ED-19, pp. 826–826, 1972.

[11] W. J. Tsa i *et al.*, "Cause of data retention loss in a nitride-based localized trapping storage flash memory cell," in *Proc. 40th Annu. Int. Rel. Phys. Symp.*, Dallas, TX, 2002, pp. 34–38.