Ultra-Fast Erase Method of SONOS Flash Memory by Instantaneous Thermal Excitation

Dae-Chul Ahn, Myeong-Lok Seol, Jae Hur, Dong-II Moon, Byung-Hyun Lee, Jin-Woo Han, Jun-Young Park, Seung-Bae Jeon, and Yang-Kyu Choi

Abstract—An ultra-fast erasing process which acts within 200 ns is demonstrated in a junctionless gate-all-around nanowire SONOS (Silicon-Oxide-Nitride-Oxide-Silicon) device. Rapid erasing is enabled with the use of instantaneous thermal excitation (TE) through a double-ended gate structure. Charges inside the silicon nitride layer are de-trapped by Joule heating. Moreover, an in-situ self-annealing effect accompanied by the TE erase method is achieved; hence, both the tunnel oxide quality and the retention characteristics are less degraded compared to the conventional Fowler-Nordheim (FN) erase method.

Index Terms—Gate-all-around, Joule heat, Junctionless, Self-annealing, SONOS, Thermal excitation.

I. INTRODUCTION

The junctionless (JL) gate-all-around (GAA) SONOS device has drawn considerable attention for the 3-dimensional (3-D) vertical integration of charge-trap flash (CTF) memory owing to its inherent immunity to the floating gate coupling issue between adjacent cells and its simple device fabrication compared to an inversion-mode GAA SONOS device [1]-[3]. Throughout the history of CTF memory, many studies have been conducted to find a way to improve the program/erase (P/E) efficiency for high-speed applications. Although a dopant-segregated Schottky-barrier (DSSB) MOSFET was suggested as the promising solution due to its high-speed program efficiency, the erase efficiency remains problematic owing to slow erase speed [4]. Furthermore, despite the fact that various technologies to enhance the erase efficiency have been proposed, e.g., 1) high-k blocking oxide with a high work-function metal gate [5], 2) a bandgap-engineered (BE) tunneling barrier [6], and 3) a BE-blocking barrier [7] structure, these CTF memory devices also show erasing times on the order of milliseconds. On the other hand, CTF memory devices are associated with reliability issues as the number of P/E cycles is increased. These issues include data retention problems caused by the aging of the tunnel oxide [8]. Moreover, the high electric field required to de-trap electrons from the nitride layer deteriorates the tunnel oxide, the condition of which is characterized by a worsened subthreshold swing (SS) due to the existence of residual interface traps. Thus, research to improve the erase speed and the reliability of the tunnel oxide is crucial.

It has been reported that during the data retention period, loss of the trapped charge distributed near the nitride conduction band [9] is rapidly accelerated as the temperature is increased due to the process of thermal excitation (TE) [10], as shown in Fig. 1(a). As a contrary concept, however, the charge de-trapping phenomenon via the TE process can be favorably used as an efficient erase method. In this letter, TE-assisted ultra-fast erasing method is proposed. The proposed method uses the Joule heat spike with the aid of a built-in heater, as shown in Fig. 1(b), as a novel erase mechanism. In addition, it is expected that the self-annealing effect arising from the Joule heat can slow down the degradation of the tunnel oxide and the post-cycle retention characteristic, which are accompanied with the Fowler-Nordheim (FN) stress. Thus, the in-situ self-annealing effect, which is not available in the traditional FN tunneling erase method, is experimentally verified.

II. EXPERIMENTAL DETAILS

The process flow of the fabricated device can be found in previous work by the authors [11]. The scanning electron microscopy (SEM) image in Fig. 2(a) shows a double-ended gate structure, i.e., a built-in heater. Cross-sectional...
transmission electron microscope (TEM) images of the fabricated device along the a-a’ and b-b’ directions are shown in Figs. 2(b), (c), and (d), indicating a gate length (L_G) of 50 nm and O/N/O thicknesses of 3 nm, 6 nm, and 8 nm, respectively. In the proposed TE-based erase method, current flow from Gate 1 to Gate 2 in the double-ended gate structure, as shown in Fig. 2(a), generates instantaneous Joule heat, which allows the trapped charges inside the nitride to be excited and finally de-trapped. Figs. 2(e) and (f) show simulated data with the aid of COMSOL [12] when voltage is applied to each side of the gate. Initially, the current through the double-ended gate is measured with respect to the erase voltage (V_{TE,ERS}) when it ranges from 0 V to 8 V. Secondly, the measured currents is plugged into the COMSOL simulator as input data with the aforementioned device structure data, as shown in Fig. 2. Thereafter, the resulting erase temperatures (T_{ERS}) in the nitride layer were extracted and plotted, as shown in Fig. 2(e). A transient temperature simulation with various applied V_{TE,ERS} values was also conducted, confirming that T_{ERS} starts to become saturated near 50 ns, which can allow ultra-fast erasing compared to the millisecond order of the magnitude in the FN tunneling erase method. Due to the current flowing between two gate electrodes, power consumption of the TE erase method is higher than that of the FN erase method. However, if low resistivity metals such as Al, W, and TiN utilized in the conventional CMOS process are used for the gate material, power consumption can be remarkably decreased, which is verified by the COMSOL simulator (data not shown).

III. RESULTS AND DISCUSSIONS

Fig. 3(a) shows the typical erase characteristics based on the TE method. After the device was programmed at a program voltage (V_{PGM}) of 14 V and a time (t_{PGM}) of 100 μs, the erase characteristics for various pulse times were investigated. To erase the programmed device, V_{TE,ERS}/2 = ± 4 V is applied to the end of each gate side, while keeping the source and drain grounded. The current flow along the gate then instantaneously provokes high-temperature Joule heating in the nitride charge trapping layer (CTL), returning a programmed state to a fresh state within an erase time (t_{ERS}) of 200 ns, as shown in Fig. 3(a). The erasing time efficiency levels for various V_{TE,ERS} values ranging from 5 V to 8 V are shown in Fig. 3(b). When the V_{TE,ERS} values of 5 V and 6 V are applied up to 1 ms, the threshold voltage (V_T) of the programed state cannot be perfectly returned to the fresh state owing to the insufficient heat. However, when the V_{TE,ERS} values larger than 6 V are applied, programed state can go back to the fresh state above 100 μs with V_{TE,ERS} of 7 V and 200 ns with V_{TE,ERS} of 8 V, respectively. Because distributed trap levels of the nitride in the energy band diagram exist, as shown in Fig. 1(a), sufficient energy arising from the Joule heat is required for the complete and rapid de-trapping of the electron. Fig. 3(c) shows a comparison of the erasing time efficiency between the proposed TE erase method and the conventional FN erase method with different FN voltages from -12 V to -15 V. Whereas the t_{ERS} value to ensure a completely fresh state is on the order of milliseconds by the conventional FN erase method, the TE erase method provides a more than 10^4-fold enhancement of the erase speed compared to former case on the order of nanoseconds with an optimized condition. Moreover,
the very high $V_{FN,ERS}$ value needed to decrease the $t_{ERS}$ to the nanosecond regime may give rise to the erase saturation effect [4] and aggravate the tunnel oxide quality. Meanwhile, as illustrated in Figs. 3(b) and 3(c), there are saturation points of the $V_T$ at the fresh state when using the TE erase method because of the re-fresh state of the CTL. However, the $V_T$ continually gets lower when using the conventional FN erase method owing to the hole injection as the $t_{ERS}$ is increased as shown in Fig. 3(c). Although there is a difference of the memory window between the TE method and FN method, injected hole can lead to an incorrect estimation of the error bit due to the initial transient $V_T$ phenomenon after the FN erase operation. [15] It is no longer a concern for the TE erase method, which does not include hole injection mechanism. Fig. 4(a) shows the dumb-mode cycling endurance characteristics of the device for different erase methods without a P/E verifying shot. Both methods show similar $V_T$ endurance levels. However, the SS values are worsened more by FN erasing than by TE erasing after iterative P/E cycling due to the degraded interface quality of the tunnel oxide, as shown in Fig. 4(b). Relatively low SS values are maintained by the TE erase method due to the in-situ self-annealing effect arising from the Joule heat during the charge de-trapping process, which can recover the degraded oxide quality. Meanwhile, there is another component which degrades the SS value in the 3-D NAND flash memory. If the geometric shape of the cross-sectional channel is closer to a rectangle rather than a circle, inhomogeneous charge injection to the CTL can occur over the effectively different device width during FN P/E operation. As a result, the parallel devices with different $V_T$ may contribute to the net current, resulting in degraded SS values. Thus, a circular shaped device is highly preferred in the state-of-the-art 3-D NAND flash memory. When using the TE erase method, re-fresh state of the CTL can be obtained during each erase step, suppressing the SS value degradation by preventing an inhomogeneous charge distribution in the CTL. The sensing window to distinguish each memory state is narrowed during the operation of the multilevel cell (MLC) as the SS value is increased. Thus, the TE method is preferred for the higher memory density supported by the MLC operation. The post-cycling retention characteristics are shown in Fig. 4(c). Degradation of a $V_T$ window is more severe in the FN erase method than in the TE erase method after a retention time of approximately $3 \times 10^5$'s under 1 k P/E cycles. On the other hand, due to the self-annealing effect spontaneously induced during the TE erase process, the interface trap-assisted charge loss is mitigated; thus, the post-cycle retention characteristics are improved. A $V_T$ window of 80 % after 10 years will be preserved after 1 k P/E under the TE mode, which is comparable to a $V_T$ window of 88 % in a fresh device. However, a $V_T$ window of scarcely 56 % is monitored after 1 k P/E in the FN mode owing to the residual interface traps which accumulate during the cycling operation.

IV. CONCLUSION

In summary, a novel erase method based on the TE mechanism was proposed and its superiority to the traditional FN erase method was experimentally confirmed. It was observed that the erase efficiency of SONOS flash memory with the TE method was more than $10^5$-fold faster compared to the traditional FN method. Furthermore, the proposed method also harnessed the in-situ self-annealing effect arising from the TE erase process to cure damage. Hence, the oxide quality was less degraded and post-cycle retention time was prolonged compared to conventional FN method.

REFERENCES