Solid-State Electronics 94 (2014) 86–90

Contents lists available at ScienceDirect

Solid-State Electronics

journal homepage: www.elsevier.com/locate/sse

Improvement of the multi-level cell performance by a soft program method in flash memory devices

Jong Kyung Park [a], Ki-Hong Lee [b], Seung Ho Pyi [b], Seok-Hee Lee [a,b], Byung Jin Cho [a,*]

[a] Dept. of Electrical Engineering, KAIST, 335 Gwahak-ro, Yuseong-gu, Daejeon 305-701, Republic of Korea
[b] SK Hynix Semiconductor Incorporated, Bubal-eup, Icheon-si, Gyeonggi-do 467-701, Republic of Korea

ARTICLE INFO

Article history:
Received 13 August 2013
Received in revised form 4 February 2014
Accepted 26 February 2014
Available online 19 March 2014

The review of this paper was arranged by Prof. A. Zaslavsky

Keywords:
Reliability
Soft program
TANOS
Blocking oxide

ABSTRACT

A soft program method is proposed for charge-trap flash (CTF) memory devices. By adding a subsequent small positive gate pulse after main Fowler–Nordheim (FN) injection programming, early charge loss is greatly reduced. The multi-level cell performance as well as the initial flat-band voltage (V_{FB}) instability can thereby be improved by removing the trapped electrons at the shallow traps in the blocking oxide layer. The proposed soft program method is a simple but very effective way to improve the fast retention property without changing the memory structure, especially for cases where the κ-value of the blocking oxide is high.

1. Introduction

As a means of overcoming the severe scaling limitations of floating gate (FG) type Flash memory devices, the development of charge trap type Flash (CTF) devices has been accelerated due to their better scalability [1–3]. However, the inferior data retention property caused by the complex nature of charge trapping has been one of the main obstacles blocking the entry of CTF devices into the mainstream of Flash memory mass production [4–6].

Recently, as alternative approaches to overcome this challenge, new programming methods have been suggested for improving the retention property in CTF devices. Lue et al. proposed the so-called 'refill method' in nonvolatile read only memory (NROM) devices [7,10]. For program operation of the NROM device, the channel hot electron (CHE) method is usually used to locally inject electrons from the channel near the source or drain junction for 2-bit/cell operation. However, since the electrons injected by the CHE method mainly occupy shallow trap levels in the nitride trapping layer, the charge loss becomes severe, especially at high temperatures [8]. To overcome this problem, Lue et al. suggested the 'refill method', which repeats the sequence of the CHE program followed by short negative pulses for Fowler–Nordheim (FN) tunneling several times to eject electrons from the shallow traps in nitride [9,10]. The refill method can modify the electron trap energy spectrum into deeper energy levels and thereby improve the retention property. However, this method is not very effective for devices using FN tunneling injection programming, which is the main programming mechanism in modern CTF devices such as TANOS (TaN/Al_2O_3/Si_3N_4/SiO_2/Si) devices [10]. This is because the electrons from FN programming initially fill the relatively deep energy levels in the nitride trapping layer. On the other hand, in TANOS devices, the high-κ blocking layer has been identified to be the one which causes early stage threshold voltage (V_T) instability and retention degradation due to charge trapping/detrapping and carrier leakage [11,12]. Moreover, the fast detrapping of bulk trapped charges during program operation in the high-κ blocking oxide can contribute to early stage charge loss by the internal electric field and thereby threatens the multi-level cell (MLC) performance [13]. In this work, we propose a new program method, referred to as the soft program method, which can reduce the fast charge loss of TANOS devices under the retention state, and thereby improve the early stage V_T instability and the MLC performance.

2. Experimental

Typical TANOS devices were fabricated on a *p*-type Si substrate. After standard gate pre-cleaning, a 4.5 nm thick thermal SiO_2 tunnel oxide was grown on the substrate, and a 6 nm thick Si_3N_4 charge-trapping layer was deposited by low-pressure chemical

* Corresponding author. Tel.: +82 42 350 3485; fax: +82 42 350 8565.
E-mail address: bjcho@kaist.edu (B.J. Cho).

http://dx.doi.org/10.1016/j.sse.2014.02.012
0038-1101/

vapor deposition (LPCVD) on the SiO_2 layer. For the blocking oxide, the experimental splits were designed for evaluating the soft program method systematically. Two different kinds of materials, Al_2O_3 and $LaAlO_x$ (with 2% La), were prepared for the blocking oxide by an atomic layer deposition (ALD) process. The 2% La-$LaAlO_x$ has a dielectric constant of ~13 [14]. The thickness of the $LaAlO_x$ and Al_2O_3 layers was 14 and 12 nm, respectively, so that the equivalent oxide thickness (EOT) of the entire gate stack is similar (11.9 and 11.7 nm). The process conditions for the sample preparation and memory performance of the two devices are described in detail in Ref. [14]. In TANOS devices, the device performance is reported to be significantly affected by the post-deposition annealing (PDA) process conditions [14,15]. In this experiment, all the samples were annealed in a N_2 ambient at 1000 °C for 30 s for fair comparison. After post-deposition annealing, a TaN metal layer was deposited by reactive sputtering and then patterned to form gate electrodes. Post-implantation annealing was conducted at 900 °C for 30 s. Forming gas annealing was performed at 420 °C for 30 min in a N_2/H_2 ambient. CTF memory cell capacitors were patterned with a gate length/width of 100 μm/100 μm for evaluating the memory performance. The FN tunneling method was used to program or erase the Flash memory cells.

3. Results and discussion

Fig. 1(a and b) shows the conventional programming method and the new soft programming method proposed in this work, respectively. In the conventional programming method, incremental step pulse programming (ISPP) is the most widely used method to maintain a tight cell V_T distribution for high reliability and multi-level cell (MLC) applications [16,17]. The ISPP method consists of several program and verification steps, and gradually increases the program voltage (V_{p0}) to reach the targeted threshold voltage of memory cells. During each program and verification cycle, the threshold voltage is stepped up and then verified to ensure that the cells are precisely programmed, as illustrated in Fig. 1(a). This iteration of the program and verification steps is maintained until the threshold voltage of the memory cell reaches the targeted verification level from the initial state. On the other hand, in the proposed soft programming method, a relatively small positive gate pulse ($V_{soft\text{-}prog}$) is added immediately after the main program pulse (V_{pf}), followed by the program verification pulse ($V_{verification}$), as shown in Fig. 1(b). By adding a consecutive small positive gate pulse after main FN injection programming, we can beforehand remove electrons in shallow traps of the high-κ blocking oxide. These electrons in shallow traps are detrapped easily and cause early charge loss in the retention state. Because the detrapping speed of electrons in shallow traps is faster than that of electrons in deep traps under a given electric field, the small positive pulse can effectively reduce early charge loss from the high-κ blocking oxides, leading to improvement of the retention property as well as the initial V_T instability. In the implementation of the soft programming method, it is not necessary to apply the small positive gate pulse at every program and verification step, as shown in Fig. 1(b). Instead, it can be more appropriate to insert the soft program pulse only after V_T reaches the critical target voltage (e.g., "target V_T minus 1 V"). This is because it is quite difficult to apply the soft-program pulse only once, since there is no certainty that ISPP and just one soft-program pulse will be enough to complete the program operation. Then, we can effectively reduce the total programming time to minimize the number of the soft-program pulse to be applied.

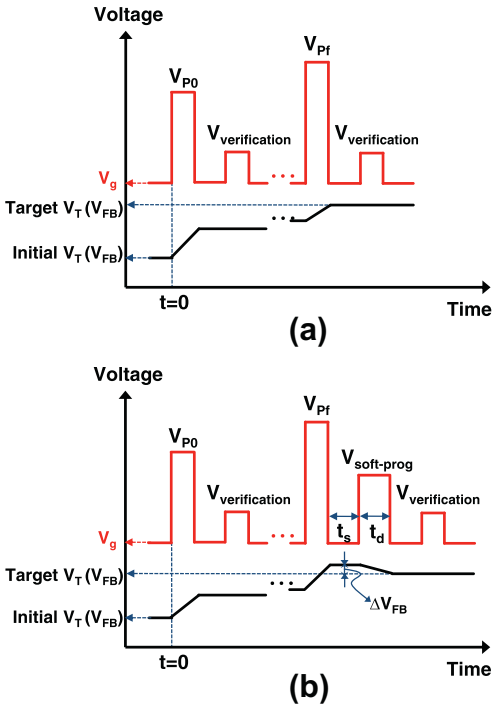

Fig. 1. (a) Conventional programming scheme based on the incremental step pulse programming method. (b) Soft-programming scheme proposed in this work: a relatively small gate pulse is applied immediately after the main program pulse. The subsequent small program pulse removes electrons trapped in high-κ blocking oxide and thus reduces the fast charge loss under a retention state.

Fig. 2(a) and (b) shows the change of the flat band voltage (V_{FB}) after program operation for different soft program voltage ($V_{soft\text{-}prog}$) and time duration, respectively. The devices were first programmed to $V_{FB} = 4$ V using a main program pulse of +18 V. The pulse duration of $V_{soft\text{-}prog}$ is fixed at 100 ms in Fig. 2(a) and the magnitude of the pulse is 8 V or 13 V in Fig 2(b). An interesting observation is the decrease of the programmed V_{FB} by applying the soft program pulse, because the positive gate pulse commonly increases V_{FB}. A more interesting phenomenon is that by adding a soft program pulse, the programmed V_{FB} decreases first and then bounces back to increase as the soft program voltage or time increases. This phenomenon can be understood by considering the different tunneling current components of the gate dielectric stack. Fig. 3 depicts energy band diagrams for three possible scenarios when the soft program pulse is applied. Here, J_D indicates the discharge current by electron detrapping from the shallow traps in the high-κ blocking oxide. J_N and J_T indicate the tunneling current from the nitride to the gate electrode and the tunneling current from the Si substrate to the nitride, respectively. When the soft program voltage is small, the electrons at the shallow traps in the high-κ blocking oxide are detrapped, whereas the electrons at deep level traps in nitride are not easily detrapped ($J_D >> J_N$). In this situation, the small change of V_{FB} can be attributed to the discharge current J_D rather than to the tunneling current J_N, since the detrapping speed of electrons at the shallow energy level in the high-κ blocking oxide is faster than that of electrons at deep level traps in the nitride by the given gate voltage. Therefore, initially trapped electrons in the blocking oxide during program operation are selectively detrapped by applying the subsequent small positive program pulse. This corresponds to region ① in Fig. 2(a), where $V_{soft\text{-}prog}$ is smaller than 8 V. As the soft program voltage further increases, the programmed V_{FB} starts to decrease abruptly (region ② in Fig. 2(a)). In this bias condition, electron detrapping from the nitride (J_N) dominates J_D because of the much higher density of trapped electrons in the nitride. On the other hand, in the blocking oxide, trapping and detrapping of electrons occur concurrently,