Solid-State Electronics 117 (2016) 123–129

Contents lists available at ScienceDirect

Solid-State Electronics

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

Understanding and optimizing the floating body retention in FDSOI UTBOX

M. Aoulaiche [b,*,1], E. Simoen [a], C. Caillat [b], L. Witters [a], K.K. Bourdelle [c], B.-Y. Nguyen [d], J. Martino [e], C. Claeys [a], P. Fazan [b], M. Jurczak [a]

[a] Imec, Kapeldreef 75, 3001 Heverlee, Belgium
[b] Micron Technology Belgium, Leuven, Belgium
[c] SOITEC, Bernin, Crolles, France
[d] SOITEC-USA, Austin, TX, USA
[e] LSI/PSI/USP, University of Sao Paulo, Sao Paulo, Brazil

ARTICLE INFO

Article history:
Available online 12 December 2015

Keywords:
1TDRAM
Floating body
FDSOI
Retention
UTBOX
Generation lifetime

ABSTRACT

The floating body retention time is investigated on fully depleted SOI devices with UTBOX. The retention is occurring through the junctions and strongly assisted by defects in the junction space charge region during the holding state at a negative gate voltage. For standard devices with a gate overlap, the junction field is high and the dominant mechanism in this case is the generation by band-to-band tunneling. For optimized extensionless devices with lower junction field, the Shockley–Read–Hall generation enhanced by the field and Poole–Frenkel mechanism takes over the band-to-band tunneling. Therefore, reducing the concentration of Si impurities closer to the junctions is the key to approach an ideal retention time only due to band-to-band tunneling with the Si bandgap as the energy barrier for tunneling.

1. Introduction

The retention time of a floating body in FDSOI devices is studied in the context of one transistor Floating Body Random Access memory (FBRAM or 1TDRAM) as a potential candidate for future Dynamic Random Access Memory (DRAM). Scaling a conventional one transistor one capacitor DRAM below 20 nm technology node is very challenging. Among these challenges, are the capacitor cell scaling and maintaining high capacitance and very low leakage, the integration scheme, the parasitic resistance increase, the sense margin reduction and the retention time degradation [1]. In one transistor FBRAM, the charge is stored in the gate to body capacitance, which circumvents all the issues related to the capacitor and its integration. Therefore, one transistor offers reduced process complexity and better compatibility with CMOS logic technology. The concept of one transistor memory was already reported in the 90s and called multistable charge-controlled memory [2]. However, it gained a lot of interest only from the beginning of 2000, when 1T-DRAM was suggested as a concept to allow manufacturing of low-cost DRAM and eDRAM below 100 nm technology node [3]. 1T-DRAM performance has been optimized targeting to satisfy the DRAM specifications for embedded and standalone memories [4–8]. Some of these specifications are fast programming, high sense margin, long retention time, and high endurance.

This paper will focus on the floating body retention time understanding and optimization. In the second part of the paper, after the introduction, we describe the device fabrication and the device operating conditions. In the third part, we will focus on the retention time analysis and discuss the crucial parameters to optimize for longer retention times. In the fourth part, we will discuss the retention time obtained by optimizing the device.

2. Experimental

NMOSFET planar devices were built on 300 mm diameter Silicon-On-Insulator (SOI) substrates with 10 nm thin BOX and 20 nm undoped Si film. After the device processing and especially for the shorter gate length devices, the BOX became thicker and the undoped Si channel thickness reduced (~18 nm-thick BOX and ~14 nm channel thickness), as measured by the Transmission Electron Microscopy (TEM). A standard 65 nm FEOL and Cu BEOL process was used for device fabrication. After STI formation a p-type ground plane doping [9] was used to adjust the threshold voltage. A gate stack consisting of 5 nm thermal SiO_2 and 5 nm

* Corresponding author. Tel.: +32 16 28 87 96.
E-mail address: marc.aoulaiche.ext@imec.be (M. Aoulaiche).
[1] Previously Imec.

http://dx.doi.org/10.1016/j.sse.2015.11.021
0038-1101/

plasma enhanced atomic layer deposition (PEALD) TiN capped with 100 nm a-Si was deposited. The thicker oxide was chosen in order to avoid the retention loss due to gate leakage current. After the gate patterning, a low-energy As-implantation was used to form junction extensions in the reference devices, while the other devices were left extension free. Next, nitride-spacers were defined using different spacer widths to vary the junction to gate overlap/underlap in the extension-less devices (15 nm, 20 nm and 30 nm). This was followed by Si-epitaxial raised Source/Drain using Selective Epitaxial Growth (SEG), and HDD implanted junctions. Among the extension-free devices, some devices received a standard highly doped junction with arsenic and phosphorus while another group of devices received in-situ doped phosphorus junctions aiming to reduce the impurities related to the junction doping and activation. A 1050 °C spike anneal was used for dopants activation. This was followed by NiPtSi silicide formation, W-filled contacts and Cu-low-k metallization.

All the devices used for this study were measured at 85 °C and have 60 nm gate length and 1 μm gate width, unless specified otherwise.

For the floating body retention measurements, an experimental setup allowing the use of short pulses of a few ns is used. Precautions were taken in order to preserve the performed waveform values and shield the signals from interferences. The fast signals were transferred through low-loss 50Ω coaxial cables with an impedance matching at the end. The cables length was optimized to avoid unwanted impedance mismatches, parasitic capacitance and crosstalk. To avoid signal bouncing and prevent ground loops all the grounds were forced to a same potential and all the probes were tied and grounded tougher close to the device under test.

The pulse width used during the charge injection or removal and during the read operation is 40 ns for both the gate and drain pulsed biases. However, during hole generation, the pulse width of the gate bias is 30 ns. The gate voltage decreased to hold prior to the drain bias in order to efficiently keep the injected holes at the front interface [10].

For the generation and the sensing of the floating body, Bipolar Junction Transistor (BJT) programming is used, so called gen2 programming [11]. To inject charges in the floating body (state 1), holes are generated by impact ionization. This impact ionization is initiated by the subthreshold current of the front channel and sustained by a BJT current once the NPN is triggered. The drain bias is set at a high value and a low gate bias is used. The generated holes are efficiently kept at the front interface by using a negative gate voltage, which creates a front channel accumulation layer. To remove the holes from the floating body (state 0), the body potential is raised by increasing the gate bias, and a small drain bias is applied in such a way that the holes are expelled via the source junction. To sense the floating body (read operation), the back channel current modulation by the presence/absence of holes in the accumulation layer at the front interface is used. If there are holes, a BJT current is triggered during the read and a high current is measured (state 1) while if there are no holes then a very low current is measured (state 0). Using this programming method, the injected holes are efficiently kept in the front channel during the hold with a negative gate voltage. The state 1 is a steady state, and a constant current is measured as a function of the holding time. However, after writing a state 0 (removing the holes) and going to the holding state at negative gate voltage, holes are generated back till reaching the steady state that is determined by the holding condition.

In this study, 50% of the steady current value measured for state 1 is used to extract the retention time as shown in Fig. 1, and the distribution of the retention times measured on transistors over different wafers is used for the analysis, as shown in Fig. 2.

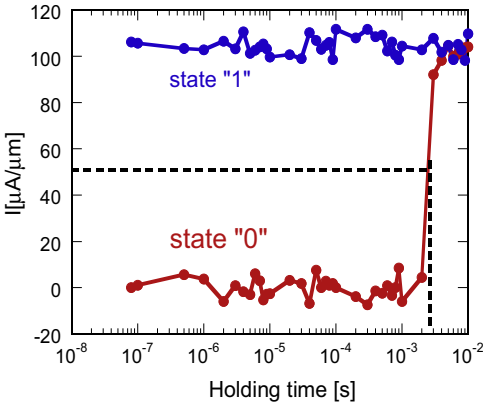

Fig. 1. State 0 and 1 current as a function of the holding time.

3. Modeling and analysis

Holding the state 0 at negative gate bias creates a demand for holes to replenish the accumulation region and go to the forced equilibrium state by the holding condition. In this case, hole supply by diffusion from the *n*+ doped junctions is negligible. Holes are generated via generation–recombination centers, at the front and back interface, in the silicon volume and they can also be generated by band-to-band tunneling between the body and junctions for a high junction field or by trap assisted tunneling at lower fields. Since we are considering BJT mode for the floating body sensing, we therefore, use the bipolar junction current expression for the read current simulation, see Eq. (1) [12]. For a negative gate bias and positive high drain bias, the accumulated holes in the front channel induce a jump in the measured drain current. This current jump is linked to the potential change in the silicon film, which forward biases the body-source junction.

$$I_{BJT}(t) = \alpha \cdot I_{F0}\left(\exp\left(\frac{q\cdot(V_S - V_{body}(t))}{K\cdot T}\right) - 1\right) + I_{R0}\left(\exp\left(\frac{q\cdot(V_D - V_{body}(t))}{K\cdot T}\right) - 1\right) \quad (1)$$

where α is the forward gain, V_S is the source potential, V_D is the drain potential, V_{body} is the transient body potential, and I_{F0} and I_{R0} are the forward and reverse saturation current, respectively. In the following, we will focus on the modeling and the understanding of the silicon body potential (V_{body}) change as a function of the holding time.

3.1. Impact of band-to-band tunneling (BBT)

Band-to-band tunneling (BBT) occurs when a negative gate bias is applied for the holding. This is due to the high electric field in the junction space charge region, especially for a standard junction device with junctions overlap. At high field, valence band electrons gain enough energy to tunnel directly to the conduction band generating electron–hole pairs. The holes flow to the body whereas the electrons go to the junctions. The BBT generation rate (G_{BBT}) is simulated using the expression (2) [13,14].

$$G_{BBT} = A\cdot F^2 \exp\left(-\frac{B}{F}\right) \quad (2)$$

where F is the local electric field, A and B are physical constants.

The potential change in the floating body is calculated using the BBT generation rate. Based on previous studies, it is assumed that the generation can occur at even 20 nm or 30 nm far from the junction [15,16]. Fig. 3 shows the simulated retention time as a