Analysis of nitride storage non-volatile memories with HfSiO$_x$ blocking dielectric and TiN metal gate for low power embedded applications

Dušan S. Golubović$^{a,*}$, Michiel J. van Duuren$^a$, Nader Akil$^a$, Antonio Arreghin$^b$, Francesco Driussi$^b$

$^a$ NXP-TSMC Research Center, Kapeldreef 75, B-3001 Leuven, Belgium
$^b$ DIEGM, University of Udine and IUNET, Via delle Scienze 208, 33100 Udine, Italy

1. Introduction

Scalability of contemporary non-volatile memories (NVMs) in commercial use to advanced CMOS nodes is a long standing issue that has led to a number of new and interesting NVM concepts, both in front- and back-end-of-line, see e.g. [1–3] and references therein. Nitride storage NVMs have been intensively studied over the past few years and may be considered, along with Si nanocrystal memories, the frontrunner to replace standard flash in the front-end-of-line [2,4].

In this paper, we investigate planar nitride charge trapping memories aimed at single bit storage embedded low power applications. In order to facilitate the low power tunnelling operation mode, hafnium silicate (HfSiO$_x$) blocking dielectric and titanium nitride (TiN) metal gate have been used. Since the TiN metal gate work function is higher compared to a standard n-flavored poly-Si gate, lower threshold voltages can be attained in the erased state due to a reduction in electron back tunnelling from the gate electrode during an erase cycle. This, in turn, improves the efficiency of the tunnelling operation with sufficiently thick tunnel oxide [1,5]. In addition to the typical NVM figures of merit, like program/erase (P/E) curves, retention and endurance, we have assessed the scalability of the devices, as well as investigated the physical origin of the observed memory features.

2. Device fabrication

The memory arrays were fabricated on 8" wafers using 193 nm deep ultraviolet lithography for active and gate definition, as well as advanced CMOS modules like pocket implants, radiation spike anneals, Ni-salicidation and so forth. The devices were integrated up to the metal-1 level using standard W- and Al-modules. In order to simplify the layout, the common source lines were implemented as buried salicided active lines. The width of the memory transistors ranges from 480 nm to 100 nm, whereas the gate lengths are between 230 nm and 80 nm.

The gate stack consists of 3 nm tunnel silicon oxide (SiO$_2$), thermally grown on a cleaned silicon (Si) surface, 6 nm silicon nitride (SiN) charge storage layer deposited by low-pressure chemical vapour deposition, 20 nm HfSiO$_x$ blocking dielectric and 10 nm TiN metal gate deposited by metal–organic chemical vapour deposition. Even though the HfSiO$_x$ layer was thermally treated in ammonia after the deposition, a high-resolution transmission electron microscope analysis (not shown here) reveals that the layer is polycrystalline. The equivalent oxide thickness (EOT) of the entire gate stack is 9.3 nm, as obtained from capacitance measurements, yielding the relative permittivity of the HfSiO$_x$ layer of around 16.

3. Program/erase characteristics

The electrical characterization has been carried out using 256 bits and two transistor (2T) 26 kb memory vehicles. The
256 bits memory array is a standard NOR array with each word- and bitline directly accessible through a separate metal bondpad. This makes it possible to apply arbitrary voltages to each word- and bitline and investigate the intrinsic properties of memory cells. The 26 kb array is a two transistor (2T) NOR array, where the source of each memory transistor is connected to the drain of a select transistor [4,6]. Thus, a memory bit can be read only if its access transistor is open, making it possible to avert over erase problems and even read out memory bits with negative threshold voltages. In this way, low voltage read of the memory, with one of the on-chip available voltages, is made possible, without the need for boosting the wordlines during read. 2T 26 kb vehicles contain on-chip drives and multiplexer, whereas the high-voltages are generated and controlled off-chip. 2T 26 kb arrays have been used for the statistically significant analysis.

The memory arrays were block programmed/erased by tunneling of carriers from the Si channel/substrate. The threshold voltages were read out using a 5 μA current criterion with 0.5 V applied to the selected bitline. In the case of 256 bits arrays, low threshold voltages in the erased state were read out by applying a slightly negative voltage to the unselected wordlines.

Fig. 1 shows the P/E curves of a 256 bits memory array with 100 nm long and wide transistors. Filled symbols show the mean value of the threshold voltage ($V_T$) in the programmed state, whereas the open symbols show the mean $V_T$ in the erased state versus the pulse width on a logarithmic scale. The average virgin $V_T$ is 1.38 V. Before each programme/erase cycle, the array was pre-erased/-programmed to fixed $V_T$ values. The use of HfSiO$_x$ in combination with the TiN metal gate significantly improves the tunnelling operation, allowing, e.g. more than 3 V $V_T$-window with ±12 V program/erase voltages.

Nevertheless, in contrast to floating gate memories where an increase of the program/erase voltage by 1 V increases/decreases the threshold voltage by 1 V, the gain in the $V_T$–window of the memory array is only around 1 V/V per decade in time. Due to a net current flow through the complete gate stack during a program or erase pulse, a part of the injected charge is lost through the gate and, consequently, does not contribute to an increase/decrease of the threshold voltage.

Fig. 2a–c shows the cumulative $V_T$ distributions of fresh, programmed and erased 256 bits memory arrays with different $W$ and $L$. All the arrays were programmed by +12 V, 1 ms and erased by -14 V, 10 ms pulses. Curves with squares are used for the arrays with $W/L$ = 0.48 μm/0.23 μm transistors, with circles for $W/L$ = 0.18 μm/0.18 μm, triangles for $W/L$ = 0.10 μm/0.10 μm and diamonds for $W/L$ = 0.10 μm/0.08 μm.
threshold voltage in the erased state is lower than the virgin value, since the latter is caused by an increased leakage current of the non-selected wordlines, whereas the erased threshold voltage decreases around 200 mV as the dimensions of memory transistors are shrunk. The latter is caused by an increased leakage current of the non-selected wordlines, since the threshold voltage in the erased state is lower than the virgin value, and a fixed current criterion is used to read out the memory. The threshold voltage in the virgin state show a very low spread, which increases as the transistor dimensions are reduced, but nevertheless does not exceed 50 mV, indicating very tight process control.

Fig. 3 shows a plot of the standard deviation of the virgin threshold voltages versus \(1/\sqrt{W/L}\), where \(W\) and \(L\) are the channel width and length, respectively. The standard deviation follows the well-known variability law modified with an additive constant \(B(Q_N)\) [7,8]:

\[
\sigma_{VT} = \frac{A_0}{\sqrt{W/L}} + B(Q_N) \tag{1}
\]

Here \(B(Q_N)\) is introduced to account for a \(V_T\) spread caused by the presence of the SiN layer in the gate stack. Namely, each fresh memory transistor has a random amount of charge stored in its SiN layer, which, in turn causes a spread of the threshold voltage even for large transistor dimensions. This phenomenon is not encountered in logic transistors with a thin gate oxide whose \(V_T\) spread very accurately extrapolates to zero as their area increases. The residual, charge induced, \(V_T\) spread of fresh memory transistors is indicated in Fig. 3 and equals \(B(Q_N) = 10\) mV. The process dependent proportionality constant is \(A_0 = 3.6\) mV/\(\mu\)m in our case.

### Table 1

<table>
<thead>
<tr>
<th>(\sigma_{VT}), (mV)</th>
<th>(W/(\mu\text{m})/(L/(\mu\text{m})))</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\sigma_{VT}^r)</td>
<td>22.5</td>
</tr>
<tr>
<td>(\sigma_{VT}^p)</td>
<td>32</td>
</tr>
<tr>
<td>(\sigma_{VT}^m)</td>
<td>52.2</td>
</tr>
</tbody>
</table>

\(W/L = 0.18\) \(\mu\)m/0.18 \(\mu\)m, with triangles for \(W/L = 0.10\) \(\mu\)m/0.10 \(\mu\)m and with diamonds for \(W/L = 0.10\) \(\mu\)m/0.08 \(\mu\)m. The standard deviations are summarized in Table 1.

The spreads in the programmed state increase around 50% compared to the virgin state, whereas in the erased state the standard deviation approximately doubles with respect to the standard deviation of the fresh arrays. As shown in Fig. 4, the relative change of the spreads in the programmed and erased state is linearly proportional to the standard deviation obtained with the fresh arrays. More importantly, both standard deviations very well extrapolate to zero implying that by improving the process control, the spreads can efficiently be kept low. The available \(V_T\)-window, measured as the difference between the programmed memory transistor with the lowest \(V_T\) and the erased memory transistor with the highest \(V_T\) changes by about 100 mV as the size of the memory transistors is reduced. This clearly implies that the trapping process is predominantly a one-dimensional process.

Two transistor 26 kb vehicles with nitride devices show the same behavior as 256 bits arrays. Fig. 5 displays the cumulative \(V_T\) distributions of two 2T 26 kb memories with \(W/L = 0.24\) \(\mu\)m/0.24 \(\mu\)m (squares) and \(W/L = 0.15\) \(\mu\)m/0.15 \(\mu\)m (circles). The arrays were programmed by 12 V, 1 ms pulses (filled symbols) and erased by -14 V, 10 ms (open symbols) or 100 ms (crossed symbols) pulses.
and erased by −14 V, 10 ms (open symbols) or 100 ms (crossed symbols) pulses.

The standard deviation for both arrays in the programmed state is around 100 mV. In the erased state, the array with \( W/L = 0.24 \mu m/0.24 \mu m \) has \( \sigma = 158.7 \) mV and \( \sigma = 141.4 \) mV for 10 ms and 100 ms erase pulses, respectively. For the 26 kb array with the smaller control gates, the standard deviations \( \sigma \) is 202.3 mV for the 10 ms erase pulse, whereas for the 100 ms erase pulse \( \sigma = 168.4 \) mV. It is important to note that in both arrays the maximum value of the erased \( V_T \), which determines the \( V_T \)-window, virtually does not change with scaling the memory transistor and the increase in the spread is caused by an increasing number of transistors with lower \( V_T \). This increases the spread in the read current, but does not impact the available memory window.

4. Endurance

Fig. 6a and b shows the endurance of 256 bits arrays for two different program/erase conditions up to \( 10^5 \) \( P/E \) cycles. In Fig. 6a +12 V, 1 ms pulses were used for programing and −14 V, 10 ms for erase, whereas +10 V 1 ms and −12 V 100 ms pulses were used for program and erase, respectively, in Fig. 6b. The vertical bars indicate the standard deviation.

The arrays do not show hard endurance failures up to \( 10^5 \) \( P/E \) cycles, but the threshold voltages in the programmed and erased states increase. Lower program and erase voltages generally cause a lower \( V_T \) increase, but are inherently associated with a lower initial \( V_T \)-window. Under all the conditions presented in Fig. 6 the array could be in principle read out after \( 10^5 \) \( P/E \) cycles, for instance, by the read voltages indicated with the dashed lines, as a sufficient difference between the programmed and erased \( V_T \) is retained. However, better endurance characteristics are required for embedded flash memories.

As the thickness of tunnel SiO2 is 3 nm, a pronounced charge trapping in it is generally not expected. Consequently, the endurance degradation can occur due to interfacial trap generation, generation of additional deep traps in the charge trapping nitride layer and/or charge trapping in the blocking dielectric, given that there is a net current flow through the gate stack during a program/erase cycle.

In order to clarify the origin of the device degradation, the mean drain current versus the gate voltage, of a fresh 256 bits array, as well as after \( 10^3 \), \( 10^4 \) and \( 10^5 \) \( P/E \) cycles with higher voltages given in Fig. 6a have been measured in the erase state. The current criteria of 3 \( \mu A \), 5 \( \mu A \) and 11 \( \mu A \) were used with 0.5 V applied to the bitlines.

Fig. 7 shows the mean drain current versus the gate voltage. The array transcharacteristic in the fresh state is given by the dashed line, whereas the number of the program/erase cycles applied before a measurement is taken is indicated next to each curve. Compared to the fresh array, the slope of the curves, that is the gain factor \( \beta \) gradually decreases and reaches a half of its initial value after \( 10^5 \) \( P/E \) cycles. This implies that interface traps are generated during cycling, contributing to a reduction of both the total capacitance of the gate stack and the channel mobility. However, the main contribution to the degradation is the trapped charge, as the curves progressively move towards higher voltages as the number of \( P/E \) cycles increases. As the \( V_T \)-window does not close, and both programmed and erased \( V_T \) increase with approximately the same rate, the charge is most likely to accumulate in the blocking dielectric [6]. It needs to be mentioned, however, that compared to floating gate flash, whose \( V_T \) evolution in the erased state is typically given by a power law \( a \cdot N^{1/2} + b \), where \( N \) is the number of cycles and \( a, b, \) and \( x > 1 \) are parameters which depend on the processing conditions, temperature and applied voltage, the erased \( V_T \) of these arrays increases as \( m = \ln(n \cdot N) \), where \( m \) and \( n > 1 \) are the processing, voltage and temperature dependent parameters [6].
5. Retention

Fig. 8 shows room temperature retention measurements of nitride devices carried out over 1 day on 26 kb memory arrays. The time evolution of the cumulative $V_T$ distributions is shown after the arrays were programmed by $+12$ V, 1 ms and erased by $-14$ V, 10 ms pulses. The arrows indicate the time. The erase voltage was chosen so as bring the erased $V_T$ close to its virgin value. For this reason the erased threshold voltage is stable over a long period of time, even at elevated temperatures, and ensures a low spread in the read current [9]. On the other hand, the programmed $V_T$ decays surprisingly fast, with approximately 200 mV/dec leading to a projected $V_T$ window, measured between the lowest programmed $V_T$ and the highest erased $V_T$ after 10 years of around 500 mV. Since the $V_T$-window decay rate is high even at room temperature, temperature accelerated retention tests have not been carried out.

Given that the tunnel SiO$_2$ with the thickness of 3 μm is at the boundary between direct and trap-assisted tunnelling mechanisms from the nitride storage layer, the initial $V_T$ window is expected to decay at room temperature, but with a lower rate than observed in this work [10]. The presence of a high-K dielectric in the gate stack causes a different electric field distribution during a program/erase cycle compared to conventional SONOS devices, as more of the applied voltage drops over the nitride trapping layer and tunnel oxide [9]. Therefore, an electron which tunnels into the conduction band of nitride becomes more accelerated by the electric field and is more likely to be injected in the blocking layer. Given the endurance curves shown in Fig. 6a and b, as well unexpectedly fast $V_T$ decay in the programmed state, charge trapping in the blocking HfSiO$_x$ layer cannot be ruled out.

By making use of the technique described in Ref. [11], we have estimated the position of the charge centroid in the gate stack in the programmed state (for details on the method we refer to Ref. [11]). It is important to note that the charge centroid position depends on the difference between the threshold voltages in the programmed and virgin states and not their nominal values. Therefore, the accuracy of the method is preserved, even though the gate leakage current is somewhat higher. Fig. 9 shows that the charge centroid is predominantly located in the blocking HfSiO$_x$ layer, even for the program pulse of 10 V, 100 μs, when the charge centroid is located almost at the interface between the SiN and HfSiO$_x$ layers.

The origin of the vertical coordinate axis is positioned at the interface between the nitride trapping layer and HfSiO$_x$ layer and increases towards the metal gate. Unlike in standard SONOS devices, where the charge centroid is located in the nitride layer for a broad range of program voltages, Fig. 9 shows that the charge centroid is predominantly located in the blocking HfSiO$_x$ layer, even if the nitride layer is at the edge of the gate stack. The charge centroid is located almost at the interface between the SIN and HfSiO$_x$ layers.

This implies that the charge is not confined to the nitride trapping layer, but a part of it is stored in the blocking HfSiO$_x$ layer for the relevant program/erase voltages. As the blocking HfSiO$_x$ layer is in the polycrystalline state, it can by no means be considered trap-free. The charge flowing through the gate stack during a program/erase cycle, given a substantial thickness of the HfSiO$_x$, is likely to be captured by the available traps in it. Note that due to a band offset difference for electrons and holes for both SiO$_2$ and HfSiO$_x$, the blocking dielectric is predominantly exposed to electron flow. This qualitative picture is entirely consistent with the observed endurance characteristics.

The lack of charge confinement in the SIN layer has a profound impact on the non-volatility of the device and is the most likely cause of the poor retention, since the charge trapped by the HfSiO$_x$ layer does contribute to the initial $V_T$ window, but can escape towards the gate much faster than the charge in the nitride layer, as it is not subjected to a confining potential.

6. Conclusion

We have fabricated and investigated nitride storage non-volatile memory arrays aimed at embedded low power applications. Uniform tunnelling of electrons/holes from the Si channel/substrate is facilitated by using a HfSiO$_x$ blocking dielectric and a TiN metal gate. Our findings on 256 bits NOR arrays have been confirmed on 26 kb memory vehicles.

We demonstrate that this type of memories can be operated by relatively low voltages, yielding, e.g. more than 3 V $V_T$-window with ±12 V program/erase voltages. It has also been shown that the spread in the programmed and erased threshold voltages is entirely governed by the $V_T$ spread of the virgin devices, and, can thus be controlled by optimizing the process integration.

It has been revealed that, even though the HfSiO$_x$ in combination with a TiN metal gate greatly improves the tunnelling opera-
tion, a considerable amount of charge is trapped in the blocking dielectric. As a result, despite the thin SiO₂ tunnel layer and tunneling operation, the threshold voltages in both programmed and erase state increase with cycling. The memories can operate up to $10^5 \text{P/E}$ cycles with no hard failures. However, charge trapping in the blocking dielectric is most likely the cause of unsatisfactory retention performance.

Acknowledgments

This work was partially funded by the Italian MIUR (FIRB RBI-P06Y5JJ project). The authors would like to thank Prof. L. Selmi for fruitful discussions.

References