Progress in Z²-FET 1T-DRAM: Retention time, writing modes, selective array operation, and dual bit storage

Jing Wan a,*, Cyrille Le Royer b, Alexander Zaslavsky c, Sorin Cristoloveanu a

a IMEP-LAHC, INP-Grenoble, MINATEC, 3 Parvis Louis Neel, BP 257, 38016 Grenoble, France
b CEA, LETI, MINATEC, F-38054 Grenoble, France
c IMEPLAHC, INP-Grenoble, MINATEC, 3 Parvis Louis Neel, BP 257, 38016 Grenoble, France

1. Introduction

The conventional one transistor–one capacitor dynamic random access memory (1T–1C DRAM) has shown good reliability and high integration density for several decades [1]. However, the external capacitor needs to store enough charge to maintain a sufficiently long retention time \( t_{\text{ret}} \), and thus is not scalable. In order to form this large capacitor, gate-all-around high aspect ratio structures have been adopted, leading to ever greater challenges in fabrication and access speed [1,2].

As a result, the capacitor-less single transistor DRAM (1T-DRAM) is of great interest due to its compact layout that dispenses with the external capacitor [3,4]. A number of demonstrated 1T-DRAMs use the floating body effect, where the stored majority carriers control the flow of minority carriers. Since the electrons and holes need to coexist in the same channel, the scaling of such floating-body memories is limited by the supercoupling effect, which impedes the simultaneous formation of accumulation and inversion layers in the same thin channel [5].

Another interesting 1T memory is based on a thyristor structure (TRAM), which shows reasonable integration density and fast access speed [6,7], but requires precise doping control to obtain stable bipolar characteristics under various temperatures [8]. A related third class of devices is based on the sharp-switching field effect diode (FED) with two front gates, which was originally proposed for electrostatic discharge (ESD) protection [9] and then as a memory device with good simulated scaling capability [9–11].

We previously demonstrated the use of a new feedback device named the Z²-FET (for zero impact ionization and zero subthreshold swing field-effect-transistor [Z²-FET]) as a capacitor-less one-transistor dynamic random access memory (1T-DRAM) through both experiment and TCAD simulation. The data retention time is measured as a function of biasing, temperature and device dimensions, leading to a simple predictive model. An alternative writing method using the source MOSFET is presented, which is potentially more compatible with the conventional DRAM array design. The operation of a Z²-FET memory array is discussed, in which the write and read signals are adapted from the single cell to achieve selective operation. Finally, we present simulations demonstrating that the Z²-FET can be used to store multiple bits thanks to the charges on both the top and bottom gate capacitors.

* Corresponding author.
E-mail address: jing.wan01@gmail.com (J. Wan).

Article info
Available online 13 March 2013
Keywords:
Z²-FET
One-transistor DRAM
Capacitor-less DRAM
Retention
Selective array addressing
Dual bit storage

Abstract
In this paper, we extend our studies on the use of zero impact ionization and zero subthreshold swing field-effect-transistor (Z²-FET) as a capacitor-less one-transistor dynamic random access memory (1T-DRAM) through both experiment and TCAD simulation. The data retention time is measured as a function of biasing, temperature and device dimensions, leading to a simple predictive model. An alternative writing method using the source MOSFET is presented, which is potentially more compatible with the conventional DRAM array design. The operation of a Z²-FET memory array is discussed, in which the write and read signals are adapted from the single cell to achieve selective operation. Finally, we present simulations demonstrating that the Z²-FET can be used to store multiple bits thanks to the charges on both the top and bottom gate capacitors.

© 2013 Elsevier Ltd. All rights reserved.
2. Experimental setup and study of retention time

The Z²-FET is a forward biased p-i-n diode with the intrinsic channel partially covered by the front gate (Lg) and the rest un gated (Lun), schematically shown in Fig. 1a. The operation of Z²-FET utilizes the feedback between electron and hole currents and the corresponding injection barriers formed by front and back gates, as described in detail elsewhere [12]. The IG–VD curves show a hysteresis window linearly controlled by the VG. To program and read, voltage pulse is applied to read out the states by inducing a transient discharging current that triggers the internal feedback (Fig. 1c). Since the |VG| decreases, prolonging the retention time (tRe) after long enough time t0 in the hold stage, it can induce a strong transient current to turn on the device and cause the failure of logic “0”, see Fig. 2b for t0 = 1.5 s. Thus, an expression for retention time can be obtained by combining the expressions of QG and ΔQG:

\[
\tau = \frac{|V_G - V_D| \cdot C_G + ΔQ_{\text{Gth}}}{I_{\text{DLEAK}}(V_G, T)}
\]

where the junction leakage |I_{\text{DLEAK}}| depends on VG and temperature T. From Eq. (1), the retention time is sensitive to the operating voltage, temperature and gate capacitance determined by the device dimensions and gate oxide.

Fig. 3 experimentally shows the dependence of tRe on VG, VD, T, LC and IS. At high |Vg|, the junction leakage is dominated by the band-to-band tunneling, which decreases exponentially as |Vg| decreases, prolonging the tRe, see Fig. 3a. At low |Vg|, tunneling becomes negligible and |I_{\text{DLEAK}}| is dominated by thermal generation, which does not depend on |Vg|, so tRe saturates. Reducing |Vg| in the readout pulse reduces the transient forward-biased voltage at the drain junction, see Fig. 3a, and hence increases tRe according to Eq. (1).

The measurements under various temperatures indicate that tRe decreases exponentially as T increases due to higher |I_{\text{DLEAK}}| as
shown in Fig. 3b. The $t_{re}$ at lower bias has stronger temperature dependence due to the thermal generation, whereas tunneling-dominated $I_{DLEAK}$ at high $|V_G|$ is less sensitive to temperature.

Reducing $L_G$ reduces the gate capacitance $C_G$ and thus reduces $t_{re}$, whereas changing $L_{IN}$ does not have a strong impact on $t_{re}$, see Fig. 3c.

3. Alternative writing mode using the source MOSFET

An alternative operation mode of the Z2-FET DRAM uses the source-side MOSFET to write to $C_G$, as illustrated in Fig. 4a. Here, $C_G$ is charged through a transistor, like a standard 1T–1C DRAM, but the stored charge is still read out through the internal feedback, ensuring less required charge and higher speed. This mode is suitable for a device with two independent gates. Here, we use the $V_{BG}$-operated Z2-FET for experimental demonstration, shown in Fig. 4b, where the $C_G$ is initially discharged through the drain junction (write “0”), and then recharged (write “1”) by the negative $V_{BG}$ pulse turning on the source-side p-channel MOSFET. We note that the charging/discharging of the gate does not require drain bias, which is important for low-power applications. In the hold state, the holes are retained on $C_G$ by the negatively biased front-gate $V_G = -1.7$ V, as in the standard Z2-FET operation above. The memory state is correctly read out with the same $V_D$ pulse via internal feedback, as discussed above. The full simulated memory cycle is shown in Fig. 4c, with constant $V_C = -1.5$ V, $V_S = -1.5$ V and 0 for discharging and charging $C_G$ as $V_{BG}$ is pulsed to -1.5 V, and both logic states correctly read out using $V_D$ pulses. It also demonstrates the access time down to 1 ns which is not achieved in our experiment due to equipment limitations. The TCAD simulation is performed in Silvaco [16] using the bipolar model with 10 ns carrier lifetime.

This mode may be advantageous because of design rules analogous to the conventional 1T–1C DRAM, with different terminals used to write and read the logic states. Another advantage of this mode is that the device stays in the OFF state during the writing of logic “1”, see Fig. 4b, consuming less power compared to the previous writing method which needs to turn on the device [13].

4. Selective read/write for memory array operation

So far, Z2-FET DRAM operation has been demonstrated experimentally on a single device only. However, any practical application of the Z2-FET DRAM requires successful array operation, with cells connected together as in Fig. 5. In the following discussion, we will assume the $V_{BG}$ of all cells is connected and biased constantly, $V_{BG} = 2$ V. The operation on a selected cell should not cause the failure or disturbance of other cells sharing the same $V_D$ and $V_S$ signal lines.

For example, the writing of a selected cell must not influence the state of other cells. Therefore, the method for writing both logic “0” and “1” to a single device shown in Fig. 2 is likely to be unsuitable, since it uses a $V_C$ pulse switching back to 0. During the writing
stage, all cells sharing the same $V_c = 0$ signal would be fully discharged and their logic states erased, as is evident from Fig. 5. Further, readout using a $V_D$ pulse only is also nonselective: in all logic "1" cells $C_C$ is fully charged and channel voltage ($V_C$) is 0 V, so the readout $V_D$ pulse forward-biases the drain junctions and turns on all logic "1" cells sharing the same $V_D$ line.

This problem can be solved by revising the logic state biasing, as illustrated in Fig. 6a. Compared to the fully charged $C_C$ in the original logic "1", see Fig. 6a, the $C_C$ is only partially charged in the modified logic "1" state, as in Fig. 6b. Suppose $V_C = -2.2$ V and voltage drop $V_{C_G}$ on the partially charged $C_C$ is $V_{C_G} = -1$ V. This leaves the channel potential $V_C = -1.2$ V, reverse-biasing the drain junction. During the readout, which still uses a negative $V_D$ pulse from 0 to $-1.2$ V, the $C_C$ is not discharged to turn on the device since the $V_C = 1.2$ V which counteracts the applied $V_D$ pulse.

In order to read out the modified logic "1" in a selected Z2-FET, the $V_G$ is first raised to $1.3$ V, which raises the $V_C$ to $0.3$ V, as shown in Fig. 6c. The subsequent negative $V_D = -1.2$ V pulse induces $0.9$ V forward bias on the drain junction and discharges the $C_C$, see Fig. 6d. The discharging current triggers the feedback and turns on the device.

Fig. 7 shows the simulation of selective operation of logic "1", where the $C_C$ is partially charged in write "1". During write "1", $V_D = -1.2$ V and $V_C = -1$ V, so that the Z2-FET is turned on due to $|V_D| > |V_C|$ with both electrons and holes injected into the channel [12,13,15]. Thereafter, $V_D$ is switched back to 0 to turn off the device and partially charge the $C_C$ due to $V_C = -1$ V. The readout using only a $V_D$ pulse down to $-1.2$ V does not turn on the device (solid lines in Fig. 7). The dashed curve in Fig. 7 shows that the device is turned on if both $V_G$ and $V_D$ pulses are used simultaneously, corresponding to a "selective" read.

The biasing for writing logic "0" can also be modified to accommodate selective operation for a DRAM array. Instead of raising the $V_G$ up to 0 and erasing all cells on the same $V_G$ line, the new method switches both the $V_C$ and $V_D$ to $-1$ V, see Fig. 8a. Meanwhile, the $V_C$ is reduced to $-1.5$ V to keep the device in OFF state. This can fully discharge the $C_C$, as indicated in Fig. 8b, without affecting other logic "1" cells sharing the same $V_C$ or $V_D$ signals. Again, the key point is that the $C_C$ is partially charged in logic "1", so $V_C = -1.2$ V and the $V_C$ and $V_D$ pulses during the write "0" sequence

![Fig. 6. Schematic view showing: (a) the original, (b) the modified logic “1” with fully and partially charged $C_C$, respectively, (c) in order to read the state, $V_C$ is first raised to $-1.3$ V, which increases the $V_C$ from $-1.2$ V to $-0.3$ V, after which (d) the negative $V_D = -1.2$ V pulse forward-biases the drain junction and triggers feedback to turn on the device.](image)

![Fig. 7. Transient simulation showing the write and read “1”. The logic “1” is read out correctly using both $V_D$ and $V_G$ signals (dashed curve). The read with only $V_D$ cannot discharge the $C_C$ and the device stays in OFF state (solid curve).](image)

![Fig. 8. (a) Simulation showing the new scheme for logic “0” operation and (b) schematic view to explain the write “0” by raising and dropping the $V_C$ and $V_D$ simultaneously to the same $-1$ V value, in order to fully discharge the $C_C$. The $V_C$ is reduced during the write “0” to keep the device in OFF state.](image)

![Fig. 9. Experimental results demonstrate that (a) the write “0” succeeds using both the $V_C$ and $V_D$ pulses, whereas the use of only (b) $V_C$ or (c) $V_D$ does not change the original logic “1” state.](image)
are not high enough separately to forward-bias the drain junction and discharge the \( C_G \) in the unselected cell.

Preliminary experimental demonstration of the selective operation is shown in Fig. 9. The device is initially written to logic “1” using the same method as in [13]. Then the \( V_G \) and \( V_D \) signals are applied simultaneously to fully discharge the \( C_G \) and write logic “0”, see Fig. 9a. Note that, in order to avoid the use of \( V_S \) (not achievable in our measurement set-up), the \( V_G \) and \( V_D \) in our experiment were pulsed to \(-0.7 \) V instead of \(-1 \) V to keep the device in OFF state. The read correctly outputs low current, since there is no charge left at the gate for triggering the feedback. Fig. 9b and c show that the write “0” with only \( V_C \) or \( V_D \) does not change the original “1” state simply because the residual gate charge is large enough to trigger the feedback and turn on the device, correctly maintaining “1” in the unselected cells.

An alternative approach for memory array operation, inspired by Section 3, is to use the source MOSFET for selective writing. This operation mode is analogous to the 1T–1C DRAM. Instead of using \( V_S = 0 \) in Fig. 4c, which fully charges the \( C_G \), the \( C_G \) is partially charged during the write “1” step by setting \( V_S = -1.2 \) V, see Fig. 10a. Then, as shown in Fig. 10b, only a combination of \( V_D \) and \( V_C \) pulses can discharge the \( C_G \) and read out the state, whereas a \( V_D \) pulse alone does not discharge the \( C_G \).

5. Scaling capability and dual bit storage

Fig. 11a shows the schematic structure of a downscaled \( Z^2 \)-FET with two non-overlapping independent gates. The simulated \( I_D-V_D \) curves under various \( V_C \) in Fig. 11b demonstrate the ultimate scaling capability of the \( Z^2 \)-FET down to \( L_G = L_{IN} = 25 \) nm, thanks to the improved electrostatic control due to ultra-thin \( T_{Si} = 5 \) nm and \( T_{BOX} = T_{ox} = 1 \) nm. The DRAM functionality is shown in Fig. 11c.
where the source MOSFET is used for writing the state, same as in Fig. 4c.

Unlike the floating body memory [4], the scaling of $T_{Si}$ in the Z$^2$-FET does not suffer from the supercoupling effect. This is due to the fact that only one type of carriers (holes in this case) is needed to trigger the feedback during the readout stage in the Z$^2$-FET DRAM, whereas most of the floating-body memories are based on the coexistence of holes and electrons in the channel.

Fig. 13. Simulations demonstrating that the dual logic bits “0,1” (a) cannot be read out through consecutive $V_D$ pulses (b) but it can be read out by alternating $V_D$ and $V_s$ pulses.

Fig. 14. Schematic view of the stored charges (left panel) and DRAM operation in simulation (right panel) of the logic (a) “0,0”, (b) “1,0”, (c) “0,1” and (d) “1,1”. Note that the initial state is “1,1” and thus there is no write step in (d).
The Z²-FET with ultra-thin buried oxide (BOX) as in Fig. 11a can be used to store two bits thanks to the large C_{BG} that provides an extra storage node. The storage capacity of the C_{BG} is demonstrated by the simulation in Fig. 12a, where the write “1” is carried out as in Fig. 11c and various V_{BG} are applied in the hold stage. For V_{BG} = +1 V, state “1” (high current) is read. As V_{BG} increases to +2 V, the read pulse outputs low current indicating that the feedback is not triggered though C_{G} is charged.

This can be explained by considering the schematic charging and discharging of both capacitors shown in Fig. 12b. During the write “1” (V_{BG} = 2 → -2 V), the C_{G} is charged by the source MOSFET, whereas the C_{BG} is discharged through the forward-biased source junction. As V_{BG} switches back to 2 V in hold stage, the channel potential in intrinsic region (V_{CIN}) is 2 V since there is no voltage dropped on C_{BG}. This deeply reverse-biases the source junction.

The negative V_{D} readout pulse (V_{D} = 0 → -1.6 V) discharges the C_{G} and enables transient hole current (I_{p}) through the drain junction. This hole current induces high electron injection from drain, which normally should flow to source and trigger the feedback. However, a large fraction of the injected electrons (I_{p}) needs to charge the C_{BG} and restore the V_{CIN}. Only the rest of electrons flow to source, forward-bias the source junction and trigger the feedback. If V_{BG} is too high, e.g. V_{BG} = 2 V, most of the injected electrons charge C_{BG} and the rest are insufficient to trigger feedback. Fig. 12c shows the evolution of the source junction bias (-V_{CIN}), where the junction is reverse-biased after write “1”. The read pulse induces electron injection and raises the source junction bias up to 1 V under V_{BG} = 1 V, which is enough to inject holes from source and trigger the feedback turning on the device. However, under V_{BG} = 2 V, the forward bias of source junction is only 0.4 V, not high enough to trigger the feedback.

Thus, the two-bit logic state in the device is “0,1”, with C_{BG} discharged and C_{G} charged. A series of consecutive V_{D} pulses cannot read out the dual logic bits, since the C_{G} would be discharged by the first V_{D} pulse, see Fig. 13a. Instead, simulations show that an alternating sequence of V_{D} and V_{S} pulses can turn on the device as shown in Fig. 13b. This is due to the fact that the first V_{D} pulse transfers electrons to C_{BG} and raises the potential of source junction to 0.4 V, as explained in Fig. 12b and c. The following V_{S} pulse (0 → 1.6 V) together with the residual 0.4 V induces high 2 V forward bias on the source junction, leading to hole injection from the source that triggers the feedback and turns on the device. Both the C_{G} and C_{BG} are fully charged by the electron–hole plasma when the device is in the ON state [13] and thus the third V_{D} pulse also outputs high current, see Fig. 13b.

Fig. 14 shows the equivalent circuits and preliminary simulations of the dual bit operation for logic states including “0,0”, “1,0”, “0,1” and “1,1”. The initial memory state is “1,1”. For logic “1,1”, since both C_{G} and C_{BG} are fully charged, the first V_{D} read pulse can induce feedback and turn on the device.

To summarize, the readout uses a 3-pulse sequence:
- Current detected at first pulse indicates “1,1” state.
- Current detected at second pulse indicates “0,1” state.
- Current detected at third pulse indicates “1,0” state.
- No current detected during the entire three-pulse sequence reflects “0,0” state.

The dual bit operation of the Z²-FET DRAM can achieve enhanced storage capacity compared to than the single bit, which translates into memory area savings. But it also offers faster writing, as only one step is needed to write two bits, see Fig. 14. However, the use of dual bit storage requires a device structure with ultra-thin BOX and non-overlapping front and back gates, the feasibility of which remains to be validated technologically on planar SOI or FinFET structures.

6. Conclusion

In this paper, we have extended our previous studies on a compact, capacitor-less DRAM device in FD-SOI utilizing the Z²-FET. The retention time of logic “0” due to the drain junction leakage current has been studied experimentally as a function of V_{C}, V_{D}, T, L_{C} and L_{BG}. The retention time increases as the voltage and temperature are reduced, due to lower leakage. The increase of gate length can also prolong the retention time thanks to the increase in C_{G}. An alternative logic state writing technique using the source MOSFET has been explored through both experiment and simulation. The selective write/read for DRAM array operation has been achieved by using partially charged gate capacitance in logic “1”. Both the selective write and read on logic “1” and “0” have been discussed and demonstrated in simulation, with preliminary experiments to confirm the TCAD results. Next, the Z²-FET DRAM was demonstrated to be scalable down to 25 nm in an advanced SOI structure with ultra-thin BOX and independent non-overlapping front and back gates. This advanced structure is capable of storing two bits in a single cell by using the charges stored on both the front and back gates and read out through alternating V_{D} and V_{S} pulses.

Acknowledgments

The work at Minatec was funded by the RTRA program of the Grenoble Nanosciences Foundation. A. Zaslavsky also acknowledges support by the U.S. National Science Foundation (award ECCS-1068895). Another author (CLR) also acknowledges support by the European STEEPER project (FP7/2007-2013, Grant Agreement No. 257267).

References


[16] Silvaco, Atlas version 2.10.4. R.