Importance of Capacitor-Less DRAM and its Scaling Perspectives

Pranjali Vatsalaya

Abstract: ***In this paper, we analyze the need for a Capacitor-less DRAM and explore the physics behind the operation modes and the scaling limits of double-gate (DG) 1T-DRAM cells. We find that this configuration allows infinitely long retention of state "1," whereas the total retention time of state "0" is limited by band-to-band tunneling (BTBT) at the source-bulk or drain-bulk junctions. Extensive and careful scaling analysis shows that short-channel effects caused by the lowering of barrier between source/drain to bulk limits longitudinal scaling, whereas BTBT limits transverse scaling. We realize that the choice of the right geometry (L, W, tox), involves a tradeoff between current amplitudes and READ sensitivity. This paper also highlights a novel Capacitor-less Double Gate Quantum Well Single Transistor DRAM approach for even better scalability and retention time. Therefore, the introduction of a "storage pocket" for holes and the engineering of the spatial distribution of holes significantly improves cell performance. Afew other unconventional Capacitor-less DRAMs are also stated.***

Keywords: Advanced Memory Techniques, Capacitorless DRAMs, Double gate DRAMs, Scaling perspectives, Quantum-well DRAM

1. Introduction

Dynamic random access memory (DRAM) is the most common kind of random access memory for mobile/personal computers and workstations. A conventional DRAM cell has a simple structure composed of one transistor and one capacitor (1T1C) per bit. The transistor acts as a switch for input and output. Unlike SRAM, DRAM is dynamic in operation; it needs to have its storage cells refreshed every few milliseconds. Because of DRAM cell structure, its size is smaller than an SRAM cell, which has 6 transistors. Thus, allowing DRAM to reach very high density. However, at sub-30nm half-pitch, conventional DRAM cells might suffer from technological scaling issues since it is harder to build a capacitor in a small cell-area with sufficient capacitance. Therefore, the scaling of 1T1C-DRAM cells has significant obstacles due to the capacitor. However, Capacitor-less DRAM can provide significant advantages to chip manufacturers as it is just a transistor on a silicon-on-insulator (SOI) wafer. This increases the amount of memory on the chip and thereby improves its performance, making the chip a lot smaller and less expensive.

2. Birth of Capacitor-less DRAM

In capacitor-less DRAM, the conventional storage capacitor was replaced by the body capacitance of the transistor. The 1TDRAM is an exceptional example of how an undesirable phenomenon, i.e., the floating-body effect of SOI technology, can be transformed into a desirable one by storing charge in the bulk of the MOSFET. It offers several potential advantages compared to conventional DRAM:

- Extremely high-density, thanks to the elimination of the additional capacitor
- Low cost of fabrication
- Excellent delay to power tradeoff due to the use of SOI technology
- Possibility of taking advantage of multi-gate architectures.

The state of charge in 1T-DRAMs is stored in the SOI substrate as excess majority charge created by impact ionization or by band-to-band tunneling (BTBT) at the bulk–drain junction induced by the relative high drain voltage. When the transistor turns on, and the high drain voltage is applied, impact ionization occurs, and electron-hole pairs are generated due to the high electric field. When excess holes exist in the floating body, the cell state is defined as "1". On the other hand, when excess holes are swept out of the floating body through the forward bias on the body - drain junction, the cell state is defined as "0". By measuring the drain current difference between Read 1 and Read 0 states of the cell, we can sense whether the holes are accumulated in the floating body. In other words, a logic state is defined by creating an excess or a shortage of the majority carriers inside the transistor's body. When a number of majority carriers are stored in the SOI, the body effect changes the transistor threshold voltage (VT) and hence its on-state drive-current. This is the primary method that the capacitor-less DRAM uses to distinguish two states.

In this paper, we focus on the working principles and scaling properties of the Double Gate (DG) Capacitor-less DRAM.

3. Double gate Capacitor-less DRAM

The 1T-DRAM cell shown in Figure1 consists of a Double Gate SOI MOSFET with the two gate contacts connected to the word lines and the source and drain electrodes connected to the bit lines.The device is fully depleted due to the short L and W, and the maximum hole concentration in bulk is around $10^{6} cm^{-3}$.[1]

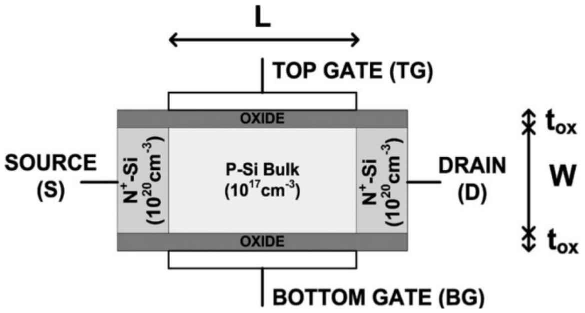
L
TOP GATE (TG)
OXIDE
SOURCE (S)
N^{+}-Si ($10^{20}cm^{-3}$)
P-Si Bulk ($10^{17}cm^{-3}$)
N^{+}-Si ($10^{20}cm^{-3}$)
DRAIN (D)
t_{ox}
W
t_{ox}
OXIDE
BOTTOM GATE (BG)

Figure 1: A 2-D DG n-MOSFET with length L=100nm, width W=100nm, oxide thickness t_{ox} = 10 nm, and bulk doping $10^{17} cm^{-3}$

WRITE Operation

During the WRITE "1" mode (W1), the two gates and the source are applied a negative bias with respect to the drain. The applied bias prevents the formation of an inversion charge at the interfaces. Electrons are injected through the source–bulk energy barrier and are collected by the relatively high potential at the drain contact. Due to the high longitudinal electric field at the bulk–drain junction, excess electron-hole pairs are created via impact ionization and BTBT processes. Excess electrons are pushed out from the bulk toward the drain due to the applied field, whereas excess holes are pushed toward the source and are trapped in the bulk if appropriate bias is applied to the electrodes.Ansignificant point to be noted is that the amount of charge representing state "1" does not depend on the actual charge generated during WRITE "1" but only on the negative top/bottom gate to source/drain bias applied during the HOLD mode. The only role of WRITE "1" is speeding up the hole generation rate in the bulk.

During the WRITE "0" mode (W0), source–bulk and drain–bulk energy barriers are lowered by applying negative potentials to the source and drain with respect to the top and bottom gates respectively. This allows removal of the charges previously stored in the WRITE "1" operation.

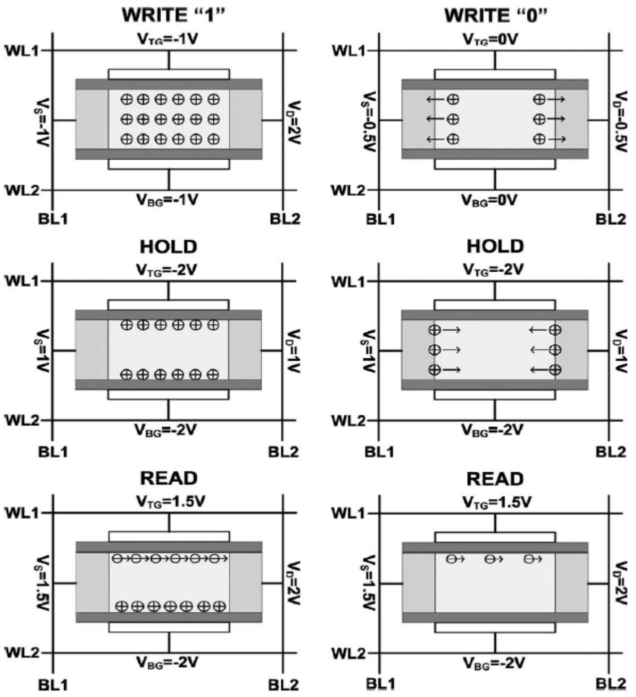

Figure 2: Bias configurations used to WRITE "1" (W1), WRITE "0" (W0), HOLD the stored data (H), and READ the stored data (R)

HOLD Operation:

During the HOLD mode (H), a negative top/bottom gate to source/drain bias is applied. This bias sets a stable accumulation condition for the two interfaces. However, if the HOLD operation follows a WRITE "0" operation, then the stable condition is reached only after a certain delay because the bulk is deeply depleted, and within this delay, the cell remains in state "0." On the other hand, if the HOLD operation follows a WRITE "1" operation, then holes are already present in the bulk because they have been generated so that the stable condition is reached quickly. Most of the excess charge generated during the WRITE "1" phase is lost by leakage through the bulk to source/drain energy barriers. The rest of the charge is accumulated at the two interfaces so that the self-consistent accumulation condition imposed by the top/bottom gate to source/drain bias is satisfied. This accumulation charge represents the information relative to state "1." Because the value of the accumulated charge is self-consistent with the applied bias, in the HOLD mode, the average hole concentration remains nearly constant for an infinite time. In other words, the retention associated with state "1" is infinity.

READ Operation:

To implement the READ operation, the two interfaces are biased asymmetrically and operated. The top interface works in a way similar to the WRITE "1" mode except that the bulk–drain reverse bias is not large enough to produce excess charge (READ disturb). On the other hand, electrons are injected from the source to the bulk and collected from the drain field. The bottom interface works as in HOLD mode (i.e., an accumulation condition) since similar potentials are applied at the source and gate electrodes. If the cell is in state "1," then the accumulation charge at the top interface is lost during READ operation because of the reduced source–bulk barrier and the accumulation charge at the bottom interface is maintained because the bottom interface is in HOLD mode.

Scaling Properties:

Since we now have a clear idea regarding the operating modes of the Double Gate Capacitor-less DRAM, let us take a look at its scaling properties. We shall explore the scaling potential of the cell as a function of its Length(L), Width(W), and oxide thickness (t_{ox}). [2]

a) Longitudinal scaling (L)

In Figure3a, we see that asL reduces, the hole concentration decreases for both states due to the increase in bulk potential induced by enhanced short-channel effects. Moreover, since the effect of the bulk potential on the hole concentration is stronger in state "1" due to the higher hole concentration, the difference in hole concentration between the two states decreases with L. It is worth noting that at sufficiently small L, this hole concentration difference disappears since the steady-state condition at the interfaces does not correspond to the accumulation regime, and our cell is unable to retain the charge at the interfaces.The READ currents for both states increases as the L reduces. This is due to the higher injection of electrons through the reduced source–bulk barrier at shorter device lengths. The ratio I_1/I_0, however, decreases rapidly due to the reduction in the difference in the hole concentration between the two states and due to the reduction in gate coupling. For sufficiently small L, the differential of the hole concentrations vanishes, and the two READ currents coincide.

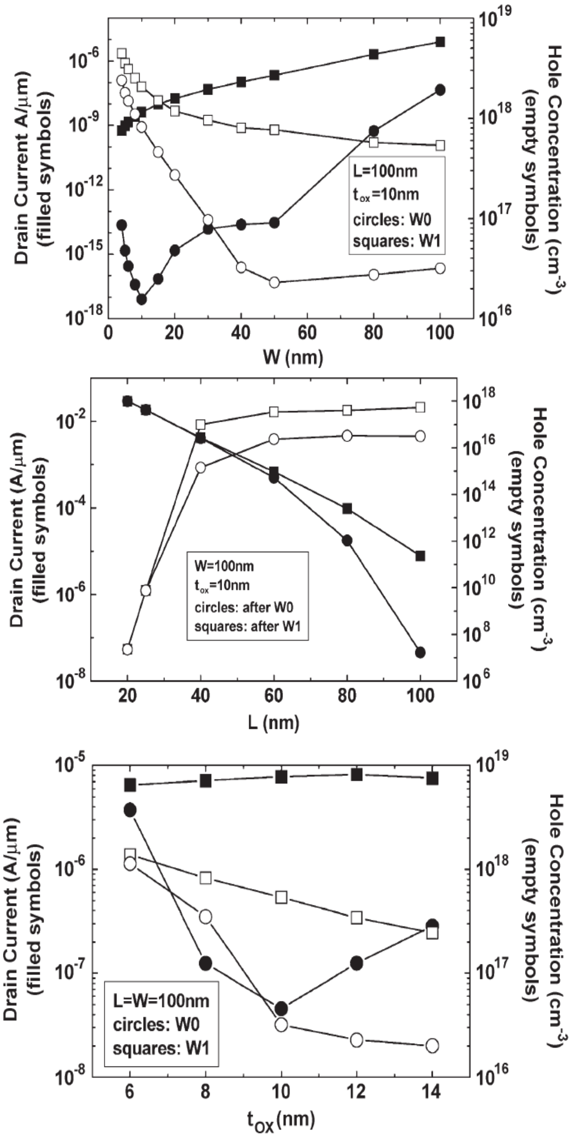

Figure 3: Drain current (filled symbols) and average hole concentration in bulk (empty symbols) after a HOLD time of 100 ms after a WRITE "1" and after a WRITE "0" a) as a function of L b) as a function of W c) as a function of t_{ox}

b) Transverse scaling (W, t_{ox})

From Figure 3b and 3c, we can conclude thatreducing W(t_{ox}) is equivalent to increasing L because it reduces SCE with the corresponding increase in the hole concentration for both states. Both currents reduce because of the lower electron injection from the source into the bulk caused by the higher source–bulk energy barrier. Moreover, in the case of W scaling, the current in state "1" decreases as well because the two gates are tightly coupled, and the charge stored at the bottom interface is lost during the READ mode due to the low source–bulk energy barrier at the top interface. As W(t_{ox}) reduces, the gate coupling increases and the ratio I_1/I_0 increases as well. When W(t_{ox}) is sufficiently low, the steady-state hole concentration in the bulk after WRITE "0" is so high that it is indistinguishable from the accumulation charge stored after WRITE "1," and as a result, the ratio I_1/I_0 starts to decrease. It is apparent that, as t_{ox} is scaled down, the minimum allowed L is reduced.

4. Capacitor-less Double Gate Quantum Well Single Transistor DRAM

For materials such as Ge and III–V material systems, passivation of interface traps is problematic and hence serves as a limitation to scaling. Therefore, we now look at a variation of DG 1T-DRAM cell, which can show higher V_T shifts and longer retention time, apart from better scalability. [3]

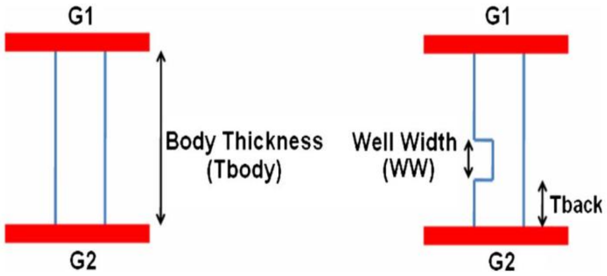

Figure 4: Schematic of (a) the band diagram of the DG Capacitor-less DRAM and (b) the band diagram of the DG hetero-structure capacitor-less DRAM (1T-QW DRAM) illustrating the QW in the body of the device.

The device is written using impact ionization near the drain. Applying a negative bias to the back gate helps to keep the holes in the created floating-body storage node. The reading is done by switching on the front MOS structure without disturbing the memory state stored in and sensing the current and determining the state by the current difference (i.e., V_T shift). In the "1" state, cells have excess holes. Thus, due to the increased body potential, they have higher drain current because of the body effect. Hence, the current difference between the two states of "1" and "0" tells the state of the device.

By introducing a "storage pocket" and by engineering the spatial distribution of the holes, it is possible to bring the stored holes closer to the front gate, thus, contributing to an increase in the amount of the V_T shift. Moreover, in many cases, keeping the stored holes away from the back oxide interface is desirable due to the presence of traps and dangling bonds which further aids in improving extrinsic retention.

5. Conclusion

In this paper, we have presented a study aimed at understanding the operation modes, the potential performance in terms of READ sensitivity, programming windows, and retention time, and the scalability of a DG 1T-DRAM cell with respect to 1T-1C DRAM cells. We noticed that the choice of longitudinal and transverse dimensions is a tradeoff between speed, READ sensitivity, retention, and programming windows. It was found that the scaling limit of device length is around 15 nm. All these significant results support 1T-OC cell as a potential replacement for classical 1T1C-DRAM. The 1T-QW DRAM has shown to have several advantages in terms of performance and scalability.Another variation in Capacitor-less DRAM is a body-tied partial-insulated FET (PiFET). PiFET structure using partially insulated oxide (PiOX) formed on bulk wafer can act as a 1T-DRAM by applying a negative back bias. The memory shows a good “0”-state retention characteristic due to a reduced electric field. The body-tied PiFET provides a wider design window and flexibility to control retention characteristics than silicon on insulator (SOI) FET. With the excellent heat immune property and good retention characteristics, a body-tied PiFET 1T-DRAM is a promising candidate for embedded memory [4]. Another 1T-DRAM architecture of interest is the Z2-FET, a band modulation based device initially developed for sharp switching, ESD protection, and memory applications. The Z2-FET is a PIN diode with gate underlap which operates in forward mode. Systematic measurements confirm that Z2-FET can deliver a significant current margin with low programming voltages, which consequently is very useful for low-power memory applications [5].

References

[1] G. Giusi, M. A. Alam, F. Crupi and S. Pierro, "Bipolar Mode Operation and Scalability of Double-Gate Capacitorless 1T-DRAM Cells," in *IEEE Transactions on Electron Devices*, vol. 57, no. 8, pp. 1743-1750, Aug. 2010, doi: 10.1109/TED.2010.2050104.

[2] N. Butt and M. A. Alam, "Scaling Limits of Capacitorless Double Gate DRAM Cell," *2006 International Conference on Simulation of Semiconductor Processes and Devices*, Monterey, CA, 2006, pp. 302-305, doi: 10.1109/SISPAD.2006.282896.

[3] M. G. Ertosun, P. Kapur and K. C. Saraswat, "A Highly Scalable Capacitorless Double Gate Quantum Well Single Transistor DRAM: 1T-QW DRAM," in *IEEE Electron Device Letters*, vol. 29, no. 12, pp. 1405-1407, Dec. 2008, doi: 10.1109/LED.2008.2007508.

[4] D. Bae, S. Kim and Y. Choi, "Low-Cost and Highly Heat Controllable Capacitorless PiFET (Partially Insulated FET) 1T DRAM for Embedded Memory," in *IEEE Transactions on Nanotechnology*, vol. 8, no. 1, pp. 100-105, Jan. 2009, doi: 10.1109/TNANO.2008.2006502.

[5] M. S. Parihar *et al.*, "Low-Power Z2-FET Capacitorless 1T-DRAM," *2017 IEEE International Memory Workshop (IMW)*, Monterey, CA, 2017, pp. 1-4, doi: 10.1109/IMW.2017.7939093.

Author Profile

Pranjali Vatsalaya is a 3rd-year student, currently pursuing B.Tech in Electrical Engineering from Indian Institute ofTechnology, Madras. Her Research interests include Analog IC Design, Advanced Memory Technology and Communication Systems. She is currently working on a project on Signal Processing and Wireless communications. Apart from being an avid reader,she holds the prestigious NTSE scholarship and has also represented her college in various inter-college Volleyball tournaments.