Retention Time Characterization and Optimization of Logic-compatible Embedded DRAM Cells

Anh Tuan Do, He Yi, Kiat Seng Yeo, and Tony T. Kim
VIRTUS, IC Design Centre of Excellence, Nanyang Technological University, Singapore

Abstract
Logic-compatible 2T and 3T embedded DRAMs (eDRAM) have recently gained their popularity in embedded applications because of their high density and good voltage margin. The most important design requirements for eDRAMs are cell area, data retention time and read speed. In this paper, we present an in-depth analysis on the data retention time of various logic-compatible eDRAM cells, followed by the effects of several design factors on the retention time. A systematic methodology is proposed for enhancing the retention time of the eDRAM cells. Simulation results using a standard 65nm CMOS technology show that the optimization process improves the data retention time more than 3×. Finally, the number of read operations per retention period is estimated to show the effectiveness of each eDRAM cell. Analysis demonstrates that although the 2T eDRAM cell has a shorter retention time than the conventional 3T cell, it has better effectiveness due to the faster read operation.

Keywords
Embedded DRAM, gain cell, retention time

1. Introduction
Low-power, high-performance multi-level caches are highly demanded to meet the ever-increasing bandwidth requirements of the core processors. Conventional SRAMs are used as the mainstream topology because of their high-speed and non-refreshing operation. However, the bulky 6-transistor cell area makes it less attractive in high-capacity cache implementations. Recently, logic-compatible embedded DRAMs (eDRAMs) have emerged as an alternative solution for complementing conventional 6T SRAMs [1-6]. eDRAM cells are much more attractive than SRAMs due to the higher density (~50% reduction in cell size when compared to a 6T SRAM cell), compatibility to mainstream CMOS processes, and non-destructive read operation. Several eDRAM cell topologies have been proposed offering high-speed and high-throughput capabilities with data retention times in the order of several hundreds of microseconds (µs) [1-6]. Zhang et. al. analyzed the performance of the 3T cell under process variations, mainly focusing on the read delay [3]. To our best knowledge, no previous work has systematically characterized the data retention time, which will be a paramount challenge when designing eDRAMs in nanoscale CMOS technologies. This paper presents an in-depth characterization and optimization of the retention time of eDRAM cells, followed by their read performance comparison.

After this introductory section, we analyze the leakage mechanism in the eDRAM cells in Section 2; followed by the retention time characterization in Section 3, Section 4 compares the performance of four designs, including both data retention time and read performance. Section 5 concludes the paper.

2. Leakage in Logic-compatible eDRAM Cells
2.1. eDRAM cells operation principles
Fig. 1 shows four eDRAM cell topologies studied in this work. All of them share the same write (M1) and storage (M2) structure. The 3T cells have one more read device (M3) while the 2T design performs its read operation by utilizing the storage device coupled with the polarity of the read control signals (i.e. RBL and RWL respectively). Their operating principles are briefly described as follows:

Conv_P [3] consists of 3 PMOS devices with the associated control signals shown in Fig. 1(a). During standby, WBL and RBL are pre-discharged to ground level while RWL is pre-charged to VDD. WWL is usually pre-charged to a voltage level higher than VDD to suppress the sub-threshold leakage current from the cell storing node to WBL. During the write operation, WBL is driven to either VDD or zero depending on the write data. At the same time, WWL is changed to a negative voltage to eliminate the threshold voltage drop across M1. A strong "1" or "0" hence can be written into the cell's storing node. The read operation is started by triggering RWL to ground, thus turning on the read device M3. Since RBL is pre-discharged low, it will be charged up to a higher voltage potential if the cell is storing a "0", i.e. M2 is on. Otherwise, if the cell stores a "1", M2 is off and hence voltage change on RBL is ignorable. The charges on RBL node will be sensed by a bit-line sense amplifier to determine the stored value of the cell.

Figure 1: Transistor level schematic of four eDRAM cells under consideration. (a) Conv_P [3] and (b) Conv_N [3] and (c) 3T_NP [4] and (d) C2T_NP [5].
The Conv_N [3] design (Fig. 1(b)) has a similar structure as the Conv_P [3] with all the PMOSs replaced by the NMOS devices. Thus, it has a better read but shorter data retention time characteristic. The C3T_NP [1] (Fig. 1(c)) uses a PMOS device for the write operation and two NMOSs for read and storage to help reduce the leakage while maintaining a good read speed. The C2T_NP [5] (Fig. 1(d)) consists of one PMOS for write operation and one NMOS for both read/storage purposes. It differs from 3T_NP [1] cell only in its read operation, whose operation details can be found in [5].

### 2.2. Leakage components and their impacts

We will investigate all of the leakage current components that flow in or out of the cell storage node and then identify the main contributors to the cell data retention. We use the Conv_P cell in Fig. 1(a) as a representative design to illustrate the leakage mechanism in eDRAM cells. Similar conclusions can be applied to the other cells with no difficulties. Since the data retention time characteristic of the eDRAM cells involves only the write and the storage devices (M1 and M2), the read device (M3) in the cell is temporarily ignored for this discussion for the sake of simplicity.

There are four leakage components that affect the data retention of the Conv_P, namely the gate leakage current of M1 \((I_{G1})\), the gate leakage current of M2 \((I_{G2})\), the body-to-source junction leakage current of M1 \((I_{J1})\) and the sub-threshold current from the storage node to the WBL \((I_{sub})\), as shown in Fig. 1(a).

Considering a cell storing "0" as shown Fig. 1(a), most of the biasing voltages surrounding the cell storage node are at \(VDD\) or higher. Therefore, the overall leakage current flows "into" the storage node and charges it up until the incoming leakage is balanced with the outgoing leakage. Fig. 2 decomposes the total "in" current into its different components (the total "out" current has only one component – the sub-threshold current through the write transistor \((I_{sub})\)). The "in" current consists of three components, among which the body-to-source reverse biased current of the write transistor \((I_{J1})\) is the dominant component. It is more than 15x the others in the 65 nm/1 V supply CMOS process used in this simulation. It is also noticeable that when the storage node voltage is about 0.3 V or higher, the total gate leakage current \((I_{G1} + I_{G2})\) becomes negligible and the body-to-source junction current \((I_{J1})\) becomes the sole contributor to the "in" current.

Fig. 2(b) illustrates the total "in" and "out" currents at the cell storage node versus the instantaneous potential of the storage node. When the storage node voltage is zero, the voltage across the drain and the source of M1 is also zero and thus no sub-threshold leakage current is observed. On the other hand, the total "in" current is at its maximum value. As the node voltage gradually increases, the total "in" current reduces while the "out" current remains almost the same. As the cell node approaches \(VDD\), the sub-threshold current through M1 increases significantly and so does the total "out" current. The steady-state node voltage is formed at a point where the total "out" and total "in" currents are equal in magnitude. If one forces the node to a higher potential, naturally, the outgoing sub-threshold current will increase while the other incoming leakage current (gate and junction currents) will decrease. Hence the storage node will discharge to the same steady-state level due to the smaller "in" current when compared to the "out" current.

We also study "in" and "out" currents of the cell at 100°C (not shown) to see the impact of temperature on leakage level, data retention time, and steady-state value. We found that the absolute differential value between the "in" and "out" current at 100°C is higher than those at 20°C and hence the rate of change of the node voltage is higher, i.e. shorter retention time. Also, the steady-state value of the cell voltage at 100°C (750 mV) is lower than that at 20°C (920 mV).

### 3. Retention Time Characterization

We investigate the impacts of six design parameters (Write transistor body and gate biasing, channel lengths and widths of the write and the storage devices) on the data retention time. The data retention time is defined as the time at which the voltage difference between the cells storing a "0" and a "1" is 250 mV, which can be sensed by our bit-line...
sense amplifier. Other values can also be chosen depending on different system requirements.

3.1. Effect of gate biasing voltage

Gate biasing voltage of the write transistor is crucial in enhancing the retention time of the eDRAM cell because it determines the sub-threshold leakage current $I_{sub1}$. Fig. 3(a) shows the data retention time of the four designs versus different gate biasing voltages of the write transistor. When the gate biasing voltage increases, the retention time is improved because the sub-threshold leakage current ($I_{sub1}$) is reduced exponentially and so do the "out" current and the rate of discharging data "1". However, this trend only extends to a certain voltage, beyond which the retention time reverses. For example in Fig. 3(a), the Conv_P, Conv_N and the C2T_NP have its retention time peaks at 1.2V gate biasing and reduces after that.

This phenomenon can be explained as follows: As the gate biasing voltage increases, the sub-threshold leakage current (i.e. the "out" current) will be suppressed and thus the node storing a "1" will take more time to discharge. A higher gate biasing voltage also shifts the crossing point of the "in" and "out" currents in Fig. 2(b) to the right and hence resulting in a higher steady state voltage, i.e. an improved retention time. However, if the biasing voltage exceeds a certain value, further reduction in the sub-threshold current becomes insignificant and its impact on the node storing a "1" is negligible. Fig. 3(b) shows the impacts of the gate biasing voltage on the data retention time and the steady-state level. As the gate biasing voltage increases from 0.9 V to 1.1 V, the steady state value of node "1" gets higher but as the biasing voltage reaches 1.1V-1.5V, the node value stays at $VDD$. Moreover, it induces a larger voltage difference along the channel below the gate, creating a stronger reversed biasing across the body-to-source junction of the write transistor. This leads to an increase in the body-to-source current which charges up the node storing a "0" with a faster rate (Fig. 3(b)), thus jeopardizing the retention time.

![Figure 3: (a) Data retention time of the four designs versus different gate biasing voltages of the write transistor. For the Conv_N cell, the corresponding gate voltage is (1-x axis value). (b) impacts of write transistor gate biasing voltage on data retention time (Conv_P).](image)

![Figure 4: Data retention time of the four designs versus different body biasing voltages of the write transistor](image)

![Figure 5: Data retention time of the four designs versus different channel length of (a) write transistor (b) storage transistor.](image)
3.2. Effect of body biasing voltage

Fig. 4 shows the data retention time of the four designs versus different body biasing voltages of the write transistor. The Conv_N cell has a NMOS write device and hence its body cannot be biased separately. Except that, the other three designs have longer retention times with a reduced body-biasing voltage. This is because when the body biasing voltage is reduced, the body-to-source current flowing to the storage node is reduced while the other currents remain the same. Thus the cell storing a “0” takes more time to charge up while the cell storing a “1” is almost unaffected, resulting in a longer retention time.

3.3. Effect of channel length

Fig. 5 illustrates the effect of the channel length on the data retention time. It is apparent that a longer channel length results in a longer retention time. However, each transistor has different impacts on the data retention time. In Fig. 5(a), the changes are less linear and smaller since a longer channel width results in a longer retention time. However, each channel width has different impacts on the data retention time. It is apparent that a longer channel width does not affect the source capacitance of the write transistor $M1$. However, it reduces the sub-threshold leakage current through $M1$ and hence the resulting improvement (22%) is not as good as that of $M2$ (33%).

Finally, in most of the above-mentioned results the Conv_P shows the best data retention time while the C3T_NP has the worst data retention time.

3.4. Effect of channel width

Channel widths of the write and the storage transistor have significant impacts on the retention time of the cells, as shown in Fig. 6. Both linearly improve the retention time because they linearly increase the parasitic capacitances associated with the cell storing node. However, increasing $W1$ also increases the sub-threshold leakage through $M1$ and hence the resulting improvement (22%) is not as good as that of $M2$ (33%).

Based on the results presented in the previous section, we propose a systematic approach to optimize the data retention time of the cell. It starts by estimating the layout area of the given cell using the minimum transistor sizes. After having the minimum layout drawing, the sizes of the storage and write transistors are gradually increased but within the area constraint. Since increasing $L1$ and $L2$ affects the speed it is advisable to know the read/write delay requirement as well. The gate/body biasing voltage of the write transistor should be chosen carefully as the higher is not necessarily the better for the retention time.

To demonstrate our optimization approach, the maximum allowable cell area is defined by the layout of the C3T_NP using minimum size transistors as it is the largest cell among the four. The final layout of the Conv_P, 3T_NP and C2T_NP are presented in Fig. 7. It is worth mentioning here that the Conv_P and Conv_N have the same layout by replacing the PMOS by the NMOS transistors.

Regarding the biasing voltage, our optimum bias voltages for the gate and the body of the write transistor is 1.3 V and 0.8 V, respectively. However, the impact of reducing the body biasing voltage to 0.8 V is negligible (less than 1% data retention improvement) therefore we suggest that it is more efficient to keep the body biasing voltage at 1V, in this particular case.
After optimization, the retention time of the Conv_P, Conv_N, 3T_NP and 2CT_NP are 769 µs, 702 µs, 320 µs and 537 µs, respectively. They are are 3×, 3×, 1.4×, 2.5× better when compared to the respective cells before optimization. Fig. 8 shows the data retention of the Conv_P using 10000-cycle Monte-Carlo simulation (using both mismatch and process variations), before and after optimization. As clearly shown, not only the retention time is extended, but also the voltage at node "1" presents less variation.

![Figure 8](image)

**Figure 8:** Data retention waveforms of the Conv_P cells using 10000-cycles Monte-Carlo simulations at 80 °C: (a) before optimization and (b) after optimization

Furthermore, for each design the longer we hold the data, the smaller the difference between "0" and "1" and thus it takes a longer time to perform a read operation. As a result, we use the effective number of reads (ENoR) as a key performance indicator which includes both data retention time and read speed and is defined as follows:

\[
ENoR = \frac{\text{Data Retention Time}}{\text{Max Read Delay}}
\]

(1)

ENoR of each design before and after optimization at 200µ and 300µs are shown in Fig. 9(a) and (b), respectively. It can be seen that after optimization, all designs exhibits a better ENoR, mainly because of two reasons: (1) the read port is enlarged and hence becomes stronger; (2) All cells have a better data retention time, i.e. slower charge/discharge of the storing node and hence better read margin. This leads to a faster read speed at the same retention time.

Another observation from Fig. 9 is that our proposed optimization has the least improvement on the C2T_NP design (1.02x). In contrast, the optimization improves the Conv_P and Conv_N designs by ~6× at 200 µs and ~13× at 300 µs retention time.

Comparing Fig. 9(a) with Fig. 9(b), it is noticeable that except for the C2T_NP, all other three designs exhibit a reduced ENoR when the retention time increases from 200 µs to 300 µs. This is because of the rapidly increased read delay, caused by the degraded input margin, coupled with the two cascaded MOS devices at the read port. Thus, with the same retention period, the C2T_NP cell offers the best ENoR and hence is more efficient in storing data.

5. Conclusion

This work presents a comprehensive study on the data retention time of eDRAM cells and proposes a systematic methodology to optimize it. By analysing the leakage components affecting the cell storage node, we have addressed the data retention mechanism in the eDRAM the cells and explored the effects of each individual design parameters on the resulting data retention time. The proposed optimization procedures enhance the retention time more than 3× while maintaining the same cell size as the reference cell. In addition, performance comparison shows that C2T_NP is the most efficient for high-performance applications since it allows larger number of read operations during data retention period.

References


