Enhanced retention time 2T embedded DRAM design

Amin Chegeni
Ph.D. in Microelectronics

Email: amin.chegeni@gmail.com

In this article, some methods to enhance the retention time of the 2T embedded DRAM based on cell transistors resizing are proposed. We show how to resize cell components to achieve the optimum result considering leakage currents, speed, and power consumption. The proposed methods are analyzed and post-layouts simulated in the 0.18um logic process to show the retention time enhancement.

Introduction: Memories are essential parts of computing systems and today's huge applications such as high-speed multi-core processors, System-on-Chip (SoC) deep machine learning, and neural networks make heavy data traffic between logic cores and off-chip RAMs, which cause the performance bottleneck. It results in an increasing demand for embedded memories.

There are several solutions to implement embedded memories such as SRAM and DRAM. Embedded DRAM (eDRAM) is designed to replace the conventional 6T SRAM to reduce memory area and power consumption. High-density 1T1C eDRAM which comprises one transistor and one trench capacitor requires a special and expensive process technology and has a drawback of the charge-destruction read operation. Another type of eDRAM is non-destructive Gain-Cell eDRAM (GC-eDRAM) including 2T, 2T1D, and 3T Cell structures where unlike 1T1C can be implemented on the conventional digital process and reduces the cost of manufacturing. Furthermore, GC-eDRAM offers dual-port functionality that makes it possible to read and write the cell simultaneously [1].

Among various architectures of non-destructive GC-eDRAM, 2T Cell is the densest structure with a cell area of half of the SRAM cell size [2]. Figure 1 shows the N-channel 2T Cell structure where WL and WBL are for 'write' and RL and RBL are for 'read' operations. In summary, the data on WBL is stored in the cell storage node capacitance during the 'write' operation by raising WL to V_{DD}. To read the data, RL goes down to 0V, and depending on the cell data, RBL drops or remains unchanged. Then, the sense amplifier connected to the RBL extracts the data and puts it on the output data bus. Note that as explained in [3], there is a limitation on RBL voltage swing. The V_{RBL} cannot drop further than $V_{DD} - 2V_t$ and it should be considered for designing 'read operation' related blocks.

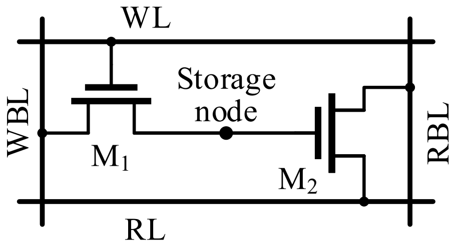

Fig 1 *2T eDRAM cell*

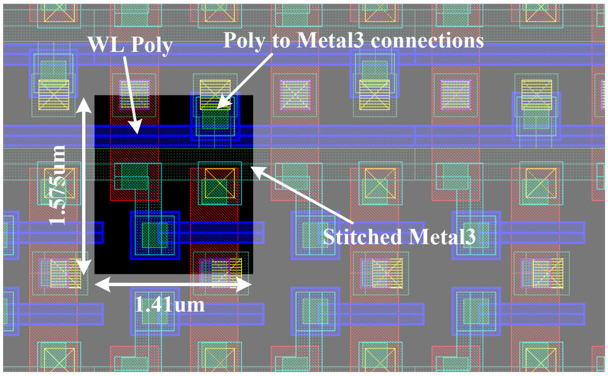

Fig 2 *Cell layout, dimensions, and Metal3 stitched on WL poly*

Aspects of the performance: There are several aspects of the GC-eDRAM performance. Array size, retention time, read/write/refresh speed, and power consumption must be considered according to their importance. They trade with each other, making the design a multi-dimensional optimization challenge. To achieve a maximum cell array density, we should choose (roughly) the minimum size of transistors M_1 and M_2. It results in minimum storage-node capacitance as depicted in (1) and hence, low retention time.

$$C_{storage} = C_{d1} + C_{g2} + C_p \qquad (1)$$

The items above represent the drain capacitance of M_1, the gate capacitance of M_2, and parasitic capacitance of the storage node respectively.

Furthermore, the smallest cell size can be achieved when WL is routed by poly [4]. Since the sheet resistance of poly is extremely greater than metal (roughly 95 times in TSMC 0.18um process), it slows down the rise and fall times of WL and therefore, the speed of eDRAM drops dramatically. So, to speed up the eDRAM, we should slightly increase the cell size to route WL by metal instead. Compared to [4], fig. 2 shows the cell layout where WL stitches to Metal 'M3'. The cell size is 1.41um × 1.575um, slightly bigger than of [4].

Considering (1), to increase the storage node capacitance, we can increase C_{g2} by enlarging the W and L of M_2 with some drawbacks. M_2 is weakened by lengthening its channel, and it slows down the read operation. On the other

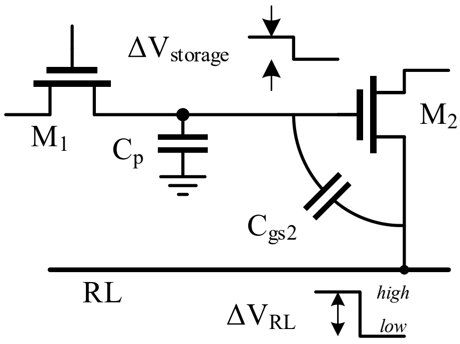

Fig 3 *Effect of VRL variation on Vstorage via M2 Gate-Source capacitance*

hand, by widening the channel of M_2, the gate-source capacitance C_{gs2} grows up and the negative effect of V_{RL} on the storage node increases as depicted in (2) and illustrated in fig. 3. It shows that when C_{gs2} increases, by lowering the V_{RL} in the read operation, the $V_{storage}$ drops more and reaches the unallowable region in a shorter time, which results in a shorter retention time.

$$\Delta V_{storage} = \frac{C_{gs2}}{C_{storage}} \Delta V_{RL} \quad (2)$$

According to (1), we can increase C_{d1} to increase $C_{storage}$ instead. However, simulation results in sub-micron processes such as 180 nm show that widening the M_1 channel not only does not increase the retention time but also slightly decreases it. The reason is that widening the M_1 channel leads to an increase of sub-threshold leakage current.

In this work, we explain that by reducing the sub-threshold current of M_1 (which extends the retention time itself), the retention time is extended through widening the M_1 channel. It is done by applying a negative voltage to the gate of M_1 to turn it off which considerably reduces the M_1 leakage current. In fig. 4 the retention time vs. W of M_1 for two values of V_{WL} are plotted. As mentioned above, for V_{wl} = 0V, the retention time is reduced slightly by increasing the M_1 width. But, when V_{wl} = – 0.25V, it grows up significantly. Fig. 5 shows the effects of V_{wl} and W of M_1 on retention time with more details. As can be seen, for a specific M_1 channel width, reducing the V_{WL} becomes ineffective after a certain value. For instance, for W of 4.2um, the value is – 0.25V and for W of 0.42um, the value is – 0.1V as shown in fig. 6. So, with a chosen width of M_1, the optimum low-level voltage of WL can be achieved.

Back to (2) and fig. 3, the variation of V_{RL} negatively affects the storage node voltage. In idle time, the RL voltage is V_{DD} to prevent M_2 from turning on. To read cell data, the corresponding RL goes down to 0V and it drops the storage node voltage via C_{gs2}. The point is, if we reduce the step-down amplitude of V_{RL}, the storage node drop-down voltage is reduced consequently and it causes the extension of retention time. Since the maximum possible value of storage node voltage is $V_{DD} – V_t$, it is enough to raise V_{RL} to $V_{DD} – 2V_t$ to turn M_2 off. However, it might

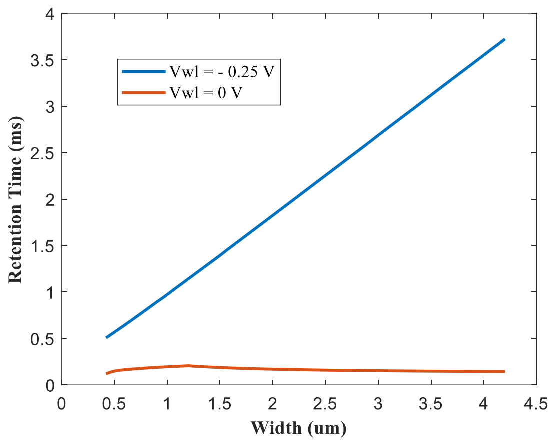

Fig 4 *Retention time vs. M1 width*

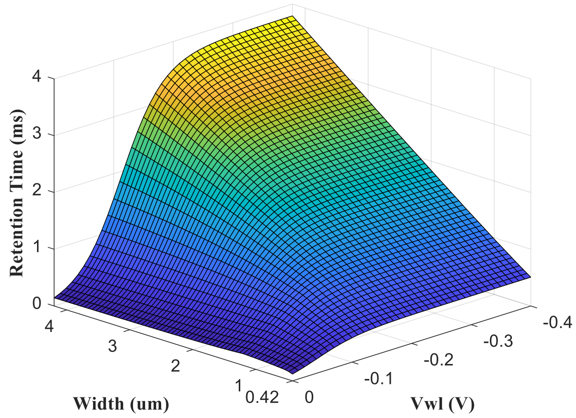

Fig. 5 *Retention time vs. M_1 width and WL low-level voltage*

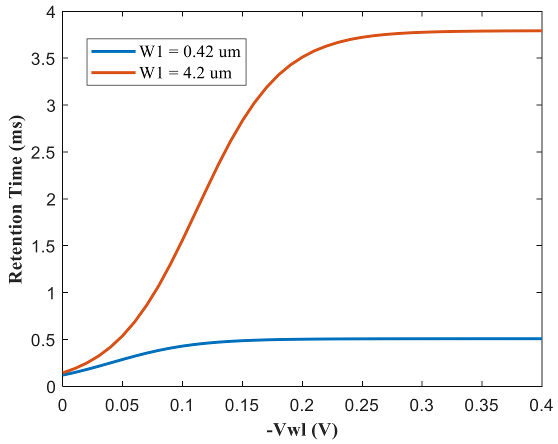

Fig. 6 *Retention time vs. WL low-level voltage for two different channel widths of M_1*

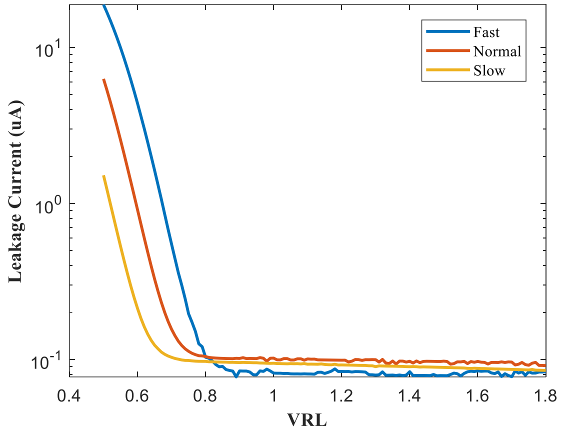

Fig. 7 *M_2 Drain-Source leakage current vs. V_{RL} in idle time*

increase the M_2 leakage current and raise the idle power consumption. So, the sub-threshold leakage current of M_2 should be considered. The maximum M_2 leakage current vs. V_{RL} in fast, normal, and slow corners for an arbitrary cell is plotted in fig. 7. It indicates that for V_{RL} above 0.8V the M_2 leakage current drops below 100nA. So, by choosing 0.8V as a high-level voltage of RL, the negative effect of RL variation is minimized with the negligible cost of power consumption. Table I shows that it extends the retention time considerably. Note that as mentioned in (2), the percentage of retention time improvement decreases for larger widths due to bigger storage node capacitances.

Table 1: Effect of reducing V_{RL} variation on retention time for various M_1 widths

M_1 Width	Retention time [a]		Improvement
	V_{RL} = 1.8V	V_{RL}[b] = 0.8V	
0.42 um	36 us	285 us	692 %
1.0 um	516 us	800 us	55%
2.0 um	1.41 ms	1.76 us	25 %
3.0 um	2.33 ms	2.7 ms	16 %
4.0 um	3.26 ms	3.65 ms	12 %

[a] for all simulations, WL low-level voltage is – 0.25V
[b] RL high-level voltage

Another advantage of reducing RL high-level voltage is the reduction of power dissipated in the RL driver. Since it is proportional to the second power of RL voltage variation, by decreasing the high-level V_{RL} from 1.8V to 0.8V, the power dissipation of the RL driver is reduced to less than a quarter.

Simulation results: The proposed 2T eDRAM is designed and post-layout simulated in TSMC 0.18um generic digital process. All retention times shown in figures 4 to 7 are extracted from simulations in all corners where the worst cases are chosen. Furthermore, 0.1V_{RMS} power supply noise is added to 1.8V V_{DD}. Note the high-level voltage of V_{RL} is 0.8V in all simulations unless otherwise noted.
Table 2 shows the enhancement of the retention time using proposed methods compared to conventional eDRAM for both minimum area and wide M_1 channel of 4um implementations. For a minimum area cell, the retention time is enhanced from 10us to 285us, and for a large area, from 130us to 3.65ms.

Table 2: Enhancement of the retention time using proposed methods

Cell area	Retention time	
	Conventional [a]	Proposed
2.22 um^2	10 us	285 us
7.86 um^2	130 us	3.65 ms

[a] WL low-level voltage is 0V, RL high-level voltage is 1.8V

conclusion: in this letter, we show that by reducing the cell leakage current and appropriate resize of the cell transistor, the 2T eDRAM retention time increases considerably without the cost of the speed and additional power consumption. To reduce the leakage current, the WL is drove to an appropriate negative voltage for logic '0' level according to the channel width of the transistor. Moreover, we decrease the logic '1' level of the RL to reduce the storage node voltage drop so that the retention time increases further. Comprehensive simulations show the enhancement of the retention time due to the proposed methods.

References

1 Kaku, M., et al.: 'An 833MHz pseudo-two-port embedded DRAM for graphics applications', *2008 IEEE International Solid-State Circuits Conference-Digest of Technical Papers. IEEE, 2008*
2 Somasekhar, D., et al.: '2 GHz 2 Mb 2T gain cell memory macro with 128 GBytes/sec bandwidth in a 65 nm logic process technology', *IEEE J. Solid-State Circuits 44.1 2008, pp. 174-185.*
3 Chegeni, A., Hadidi, K., and Khoei, A.: 'Design of a High Speed, Low Latency and Low Power Consumption DRAM Using two-transistor Cell Structure', *14th IEEE International Conference on Electronics, Circuits and Systems 2007 Dec 11, pp. 1167-1170*
4 Harel, O., Nachum, Y., and Giterman, R.: 'Replica Bit-Line Technique for Internal Refresh in Logic-Compatible Gain-Cell Embedded DRAM', *Microelectronics Journal, 2020 Jul, 101:104781*