A low leakage 500MHz 2T embedded dynamic memory with integrated semi-transparent refresh

Anselme Vignon, Stefan Cosemans, Wim Dehaene

K.U. Leuven

ESAT - MICAS Laboratory

Kasteelpark Arenberg 10, Leuven, Belgium

anselme.vignon@esat.kuleuven.be

Abstract— This paper presents a low-leakage 128 kbit dynamic memory based on a 2T dynamic cell. The design is implemented in a logic 90 nm technology and achieves a low static power consumption of 130 μ W and an access time of 2 ns. It has a worst case retention time of 175 μ s. This performance is achieved by introducing an optimized hierarchical organization and peripheral circuits for the read, the write and the refresh operations. A novel writing mechanism for 2T cells using a two phases approach is demonstrated. The area penalty of using short read bitlines is alleviated using a charge transfer sense amplifier. A novel local write sense amplifier that can operate as a latch makes it possible to perform the refresh operation at the local level, improving the energy efficiency of the refresh operation.

The memory includes an integrated automatic refresh mechanism. During refresh cycles, most read and write operations can still be performed. In cases where the address conflicts with the refresh operation, the memory handles access recovery internally,

I. INTRODUCTION

The emergence of new applications for high speed and low power devices requires more and more versatility from the memory design. Most of these applications are enabled by using memories, that are not optimized only for a speed vs density trade-off, but also for a minimum static power consumption. This is due to restrained power budget in these cases. Reaching lower power consumption for similar speed becomes a requirement in these designs. This type of low power driven design can be found in many memory implementations ([1], [2], [3], [4])

In this context, DRAM benefits from a renewed interest, partly due to the emergence of 3D interconnects, which allows for a reduced power consumption when using a standalone DRAM [5]. High speed embedded DRAM (eDRAM) macros have been demonstrated by [6]. In both cases, it has been shown that with recent technology improvement it was possible to nudge the DRAM design space to higher speed and lower power consumption, at the price of a reduced density, when compared with conventional DRAM macros.

It has been shown in [7] that DRAM memories can lead to lower static energy than SRAM. This is because DRAM cell leakage is significantly smaller than that of an SRAM cell. The challenge is to obtain a low power consumption for the refresh cycles that are needed when using DRAM.

The goal of this work is to design a DRAM memory consuming less energy than an SRAM at the same moderate

Fig. 1. Matrix organisation.

speed (500MHz). This is achieved by, on one hand, reaching a similar dynamic energy as an SRAM matrix, while on the other hand designing a power efficient refresh mechanism. This allows the power consumed by the DRAM refresh cycles to be smaller than the power consumed by the static leakage from the equivalent SRAM cells.

This work presents an implementation of such low static power matrix. We used a 2T capacitive cell inspired by [8], where it was proposed for high bandwidth applications. While both designs use a similar cell, this work is an alternative to moderate speed low power SRAM leading to a drastically different global architecture (section II), bitline structure and behavior (section III,IV), as well as refresh operation (section V). It will be shown in section VI that a moderately high speed (2ns access time), 90nm 128kbits memory matrix was designed, featuring a semi transparent refresh, 2T dynamic cells showing a 175 μ s retention time, as well as a significant improvement in static energy consumption. In low power mode, further static power reduction is achieved.

II. GLOBAL ARCHITECTURE

In order to achieve a low dynamic energy as well as a low access time, a fine granularity matrix was used [7]. The address

(a)

(b) Simulated read '0' (left) and read '1' (right).

(c) Simulated write '0' (left) and write '1' (right) access

Fig. 2. (a) Schematic of a bitline, with read sense amplifier and write sense amplifier (WSA). Simulated read (b) and write (c) access waveforms.

space consists of 4096 words of 32 bits, as described in Fig. 1, for a total of 128kbits. It is divided in 8 columns. Each column is one word wide. In the horizontal direction, the column vertical bitlines are muxed to 32 global sense amplifiers, that are shared by the entire memory.

A column is divided in 8 localblocks of 64 words that share the same write sense amplifiers (8 localblocks per column). Each of these localblocks is divided in 4 blocks of 16 words that use the same read sense amplifiers.

The 2T cell (Fig. 2(a)) consists of two transistors, connected by four terminals: the write wordline WWL , the write bitline WBL , the read wordline RWL , and the read bitline RBL . In this architecture, WBL is connected to a buffered output of the write sense amplifier. RBL is connected to the read sense amplifier. Both WWL and RWL timing are local control signals, whereas a static global wordline GWL controls which word is being accessed.

Fig. 3. Measured cumulative density distribution of the retention time for a single cell.

III. WRITE BITLINE OPERATION

The proposed bitline (Fig. 2(a)) was designed in order to maximise the 2T cell retention time (Fig. 3), to reduce the passive power.

The retention time of a DRAM cell depends on the ΔV between a cell written with a logic '1' and a cell with a logic '0'. ΔV is limited by the threshold of the access transistor, that needs to be high, in order to limit the cell leakage, thus degrading the retention time. This problem is usually solved by overdriving WWL during write, to increase the charge being stored.

However, overdriving WWL usually requires an additional voltage source, as well as a dedicated access transistor. In this design, an alternative solution for this problem was implemented, by using a two phase approach during write (Fig. 2(c)). In the first phase, a regular write is performed, with WWL at 1.2V and WBL at the value to write. During this phase, RWL is driven low and RBL is floating. During the second phase, RWL is driven at 1.2V while the access transistor remains open. The resulting voltage increase on the storage node is absorbed by the WBL if it is at 0V. If WBL is at V_{dd} , on the other hand, the storage node voltage increases, therefore increasing ΔV , when WWL is closed with RWL kept at 1.2V.

IV. READ BITLINE OPERATION

Reading a bit uses a charge transfer sense amplifier [9], that allows for a reduced area penalty when using short read bitlines.

A reading (Fig. 2(c)) or a refreshing access starts by pulling RWL low, therefore developing a voltage difference in RBL , that is resolved by the read sense amplifier, by lowering V_{bias} from 1.2V to a lower reference voltage V_{ref} . If at that time RBL is still at precharge level, the charge at RBL will transfer to the drain of the sensing transistor, causing $WRBL$ to be discharged to 0V. If, on the other hand, RBL is at a voltage less than approximately $V_{Tp} + V_{ref}$, where V_{Tp} is the threshold voltage of the sensing transistor, no charge transfer occurs.

If the memory is being read, a $WRBL$ discharge causes the global bitline $VRBL$ to be charged. In case of refreshing, $VRBL$ is gated, while $WRBL$ is used as input to the write

Fig. 4. Estimated ratio of cycles lost due to conflicts between the memory accesses and a refresh operation. In approach 1, the refresh blocks the entire memory, and a single word is refreshed per refresh cycle. In approach 2 and 3, the refresh is semi-transparent: only accesses to addresses on the localblock that is being refreshed are blocked.

sensing amplifier. By connecting the two sensing nodes of this sense amplifier to the ground, the sense amplifier is effectively transformed into a transparent latch, that is used to store the cell value, then restore this value in the cell, using the same technique as described for a writing cycle. This effectively results in a local read followed by a local write back.

Due to the other cells connected to the $\overline{RBL}$, with their $\overline{RWL}$ at 1.2V, the ΔV developed on $\overline{RBL}$ during read must be lower than the threshold voltage of the storing transistor, to avoid short circuit current. In this case, using the read sense amplifier described above allows for choosing a reference voltage, as compared with a single transistor sense amplifier. It is, on the other hand, more compact than a differential sense amplifier using a voltage reference.

V. REFRESH OPERATION

In any dynamic memory matrix at least the refreshed bitline becomes inaccessible during a refresh cycle. For this reason, it would be complex and costly to build a dynamic memory, that is fully accessible during refresh. On the other hand, having to handle failed or delayed accesses generates a cost at the system level. This cost increases with the probability of having to deal with such abnormal accesses.

A second important parameter for refreshing is the active energy consumed during the cycle, as it affects the static energy consumed by the memory [7]. Therefore, special care was taken to reduce the energy consumption during refresh.

In this work, the refresh accesses are scheduled internally from a refresh clock, that is asynchronous with the access clock. Every access is accepted, but there is a possibility for the access to be delayed in case of refresh conflict, in which case an *accessDelayed* flag is raised until the access is processed. To achieve this result, both refresh and access controls are generated by two different timing circuits, handled by a priority mechanism, that arbitrates between access, refresh, or both cycles.

This refresh mechanism blocks accesses to every address from the writeblock of the word being refreshed, while the

Fig. 5. Measured number of bits in the memory that do not achieve the retention time.

rest of the memory is still accessible. This was done in order to reduce the probability of having to block an access during refresh, as explained in Fig. 4.

To reduce the energy consumed by refreshing, but also to increase the matrix availability during a refresh cycle, multiple refresh accesses are performed during the same cycle. During each refresh cycles, one address for every columns gets refreshed (Fig. 4).

VI. MEASUREMENT RESULTS

The proposed architecture was implemented in a 90nm, logic low power multi-threshold technology. A performance summary can be found in Table I. The worst case retention time, as measured on five different dies is 175 μs (Fig. 5). This allows for a really aggressive static energy figure, of 130 μW in low power (LP) mode.

The retention time is measured as the minimal period between either cycles (read, write or refresh), when every cell is functional in a given die (Fig. 5). The lowest static energy figure of 130 μW is obtained by reducing the global decoder and timing control of the memory to the supply of the global write bitlines (0.75 V). In this mode, the memory is still functional, with a clock frequency of 250 MHz.

TABLE I
PERFORMANCE SUMMARY

Memory size	128kbits
Word size	32bits
Technology node	90nm
Cell area	0.79 μm^2
Access time	2ns
Static power	130 μW
Dynamic read energy	3.7 pJ/word
Dynamic write energy	5.2 pJ/word
Dynamic refresh energy	20.4 pJ/16 words

The static energy partitioning is shown in Fig. 6. For this architecture, the power due to the periodic refresh cycle is not the dominant energy cost.

We compared the measured results with two other published memories in Table II: the first one is an implementation of a slightly faster memory in the same technology, that also aims at reducing static power consumption [10]. The second is

Fig. 6. Static power consumption in high speed (HS) and low power (LP) modes.

another 2T implementation, aimed at reaching higher density. In both cases, this work shows a lower static power.

The density figure is higher than the one for the equivalent SRAM. However, the bit density is significantly reduced, when compared with the other 2T cells. It is difficult to draw significant conclusions from this figure, as several factors come into explaining this difference. The minimal feature size of [8] is lower, and part of the difference can be explained by the area increase between a cell layout optimized for lithography, and a cell following logic based layout rules. We did not have access to lithography optimized cell for this layout. Yet it is clear that, other things being equal, high granularity DRAM matrix have a lower density than more standard layout, as this type of architecture is focused on energy rather than on density. For the same energy/speed trade off, however, the density comparison with [10] is to the advantage of this work. To conclude, the dynamic power for this implementation is similar, when compared with other SRAM memories of the same speed.

TABLE II

COMPARISON BETWEEN THIS WORK, AN SRAM WITH SIMILAR LATENCY AND ACTIVE ENERGY, AND THE REFERENCE 2TDRAM [8]. THE STATIC ENERGY IS THE BEST AVAILABLE AT 500MHz

	This work (90nm)	[10] (90nm)	[8] (65nm)
Size	128kbits	128kbits	2Mbits
Minimal cycle time	2ns	1.2ns	2ns
Retention time	175 μ s	N.A.	10 μ s
Static power	173 μ W	457 μ W	0.768W
Static power per bit	1.35nW/bit	3.57nW/bit	384nW/bit
Area per bit	4.04 μ m ²	4.9 μ m ²	1.09 μ m ²

VII. CONCLUSION

We demonstrated a dynamic memory, that shows lower static power figure than the equivalent SRAM matrix, for the same access time, size, and technological node. This shows, that it is possible to use dynamic cells for high speed and low power memories.

Several techniques were designed to accommodate the use of 2T cells in this context. Shorter read bitlines and two phases writing were used to bring the retention time to 175 μ s.

While the area needed for the peripherals allowing such performance makes this matrix less dense than typical eDRAM

macros, it should be noted that the density is still higher than

Fig. 7. Chip microphotograph.

its SRAM counterpart. It has been shown that the refresh can be handled internally by the memory, requiring only a dedicated refresh clock from the system. Moreover it has been proven, that the occurrence of having an access delayed can be as low as 6.5% even in the worst case of a permanently active memory. This probability is less than 1% for more realistic activity scenarios.

It has been demonstrated that it is possible to reduce static power consumption for medium speed, low power memories, by using capacitance based memory cells. An implementation of such memory has been described, that shows a lower static power, when compared to an equivalent implementation using SRAM cells. A static leakage of 1.35 nW/bit for a 2ns access time is achieved by this implementation.

REFERENCES

- [1] S. Cosemans, W. Dehaene, and F. Catthoor, "A 3.6pJ/access 480MHz, 128Kbit on-chip SRAM with 850MHz boost mode in 90nm CMOS with tunable sense amplifiers to cope with variability," *Solid-State Circuits Conference, European*, Sept. 2008.
- [2] N. Verma and A. Chandrakasan, "A 65nm 8T Sub-Vt SRAM employing sense-amplifier redundancy," *IEEE International Solid-State Circuits Conference, Digest of Technical Papers*, pp. 328–606, 11–15 Feb. 2007.
- [3] S. Jain and P. Agarwal, "A low leakage and SNM free SRAM cell design in deep sub micron CMOS technology," *VLSI Design, 19th International Conference on*, p. 4, 3–7 Jan. 2006.
- [4] Noguchi *et al.*, "A 10T non-precharge two-port SRAM for 74% power reduction in video processing," *VLSI, IEEE Computer Society Annual Symposium on*, pp. 107–112, 9–11 March 2007.
- [5] M. Kawano, "A 3D packaging technology for high-density stacked DRAM," *VLSI Technology, International Symposium on*, pp. 1–2, 23–25 April 2007.
- [6] J. Barth *et al.*, "A 500MHz random cycle 1.5ns-latency, SOI embedded DRAM macro featuring a 3T micro sense amplifier," *Solid-State Circuits Conference, IEEE International*, pp. 486–617, 11–15 Feb. 2007.
- [7] A. Vignon, S. Cosemans, W. Dehaene, P. Marchal, and M. Facchini, "A novel DRAM architecture as a low leakage alternative for SRAM caches in a 3D interconnect context," in *Design, Automation Test in Europe Conference Exhibition*, april. 2009, pp. 929–933.
- [8] D. Somasekhar *et al.*, "2GHz 2Mb 2T gain-cell memory macro with 128GB/s bandwidth in a 65nm logic process," in *Solid-State Circuits Conference, Digest of Technical Papers. IEEE International*, Feb. 2008, pp. 274–613.
- [9] L. Heller, D. Spampinato, and Y. Yao, "High sensitivity charge-transfer sense amplifier," *Solid-State Circuits, IEEE Journal of*, vol. 11, no. 5, pp. 596–601, oct 1976.
- [10] S. Cosemans, W. Dehaene, and F. Catthoor, "A low power embedded SRAM for wireless applications," *Solid-State Circuits Conference, European*, pp. 291–294, Sept. 2006.