Tunnel FET Based Refresh-Free-DRAM

Navneet Gupta1,2, Adam Makosiej2, Andrei Vladimirescu1, Amara Amara1, Costin Anghel1
1 MINARC Laboratory, Institut Superior d’Electronique de Paris (ISEP) France,
2 LETI, Commissariat à l’Energie Atomique et aux Energies Alternatives (CEA-LETI) France
(costin.anghel@isep.fr, navneet.gupta@isep.fr)

Abstract—A refresh free and scalable ultimate DRAM (uDRAM) bitcell and architecture is proposed for embedded application. uDRAM 1T1C bitcell is designed using access Tunnel FETs. Proposed design is able to store the data statically during retention eliminating the need for refresh. This is achieved using negative differential resistance property of TFETs and storage capacitor leakage. uDRAM allows scaling of storage capacitor by 87% and 80% in comparison to DDR and eDRAMs, respectively. Bitcell area of 0.0275μm² is achieved in 28nm FDSOI-CMOS and is scalable further with technology shrink. Estimated throughput gain is 3.8% to 18% in comparison to CMOS DRAMs by refresh removal.

Keywords—Tunnel FET; DRAM; eDRAM; Metal-Insulator-Metal (MIM) Capacitors;

I. INTRODUCTION

CMOS technology scaling is the key factor in addressing the demand of ever increasing complexity of VLSI designs adding more and more computation power on a die. In order to meet the overall performance requirements while scaling down the devices, high throughput memory technologies are becoming increasingly important. Conventionally, Systems-on-Chip (SoCs) rely on SRAMs in order to address the throughput/performance gap between CPU and main memory (DRAM). However, SRAM’s size and power consumption is of critical concern with the rapid growth in capacity requirements of high bandwidth memories. SRAMs are used as primary cache memory in almost all kind of processing systems. Area overhead due to poor array density and leakage power consumption of high-throughput SRAMs is limiting factor in reduction of the silicon footprint and system cost. In [1], 37.5MB SRAM cache consumes more than 25% die area of Intel Ivytown Xeon processor implementation in 22nm technology. In order to further reduce the die area and cost, DRAM (DRAM) has been explored as an alternative option. DRAMs are better in comparison to SRAMs in memory density and overall throughput [2-5]. In [5], authors have reported 20% to 75% performance gain for various applications by using 1Gb eDRAM as L4 cache. However, in order to scale DRAMs aggressively specific technology and process is used, e.g. vertical transistors and capacitors. This makes the inclusion of DRAMs in logic chips difficult and costly. An intermediate solution which has been explored is the use of embedded DRAMs (eDRAMs) which are denser then SRAMs and relatively easier to fabricate with CMOS technology for digital logic technology [5-7]. In [5], eDRAM using 22nm logic technology process is used which provides higher density with 3x array efficiency in comparison to low voltage SRAMs in same technology. However, standard 1T1C DRAM structure is becoming ever more difficult to scale with technology, specifically because of difficulties in scaling the capacitance. Capacitors with high capacity are required in order to reduce the throughput penalty because of refresh requirements. This limits the scaling of DRAM capacitors. It can be noted that for the year 2016, ITRS roadmap [8] shows that the capacitance requirement per bit for DRAMs is reduced by 20% in comparison to year 2009. However, the transistor technology is scaled from 52nm to 22nm, i.e. by 57% in the duration from year 2009 to 2016. Various techniques, like negative wordline and high gate oxide thickness of capacitors, are used in DRAMs to reduce leakage and thus to increase retention time. eDRAM capacitors in [2] is using effective oxide thickness (EOT) of 0.7nm to get 8fF/bit capacitance with 0.1fA/bit leakage. However, the EOT of 0.3nm, suggested by ITRS for DRAM capacitors [8], results in significantly increased capacitor leakage. In eDRAMs capacitor size is reduced at the cost of retention time in order to optimize cost of process and silicon footprint. In [5], 14.2fF/bit capacitance is implemented in eDRAM with planar transistor achieving 22.1Mbits/mm² array density, providing only 100μs retention time while using negative WL to reduce transistor leakage. The obtained refresh power is at 1.5W/Gbit which is 30% of the eDRAMs peak active power consumption. Another critical issue, specially for eDRAMs, is that the leakage increases significantly at high temperatures which is often the case for eDRAMs because of close proximity to compute intensive blocks like CPUs/GPUs.

For example in JEDEC DDR specifications [9], the refresh time interval (RFEI) is reduced by 50% for operation above 85°C due to increased leakage in bitcells. The impact of refresh on throughput is 3.8% to 8% in best-case, i.e. assuming refresh is not blocking any read/write access. In actual scenario, due to read/write traffic interruption because of refresh commands, throughput penalty can reach up-to 12% and 18% for normal and high temperatures of operation, respectively.

In order to address the aforementioned DRAM design challenges, other than CMOS technologies have been explored. The Tunnel Field Effect Transistor (TFET) was proposed as a possible solution to reduce leakage while having scalability as MOSFETs. The TFET operates by band-to band tunneling and therefore the subthreshold slope (S) is not limited to 60mV/dec as in the case of CMOS [10, 11]. Fabricated TFETs with S as low as 30mV/dec have already been measured [12]. Progress on TFET devices has encouraged research on TFET circuits. Few reports in the literature on TFET circuits describe mostly...
the design of TFET SRAM cells [13-18]. Moreover, unidirectional behavior and negative differential resistance (NDR) properties are very promising to design circuits while addressing the issues of conventional CMOS circuits and architectures. In [17,18], ultra low leakage and compact TFET SRAM cells are proposed using NDR property of TFETs. Because of unidirectional behavior of TFETs, conventional 1T1C DRAM architecture cannot work in a similar way as for CMOS. Therefore, there is a need to optimize circuits specifically for TFETs in order to utilize different then CMOS properties of TFETs.

In this paper, refresh-free and scalable ultimate-DRAM (uDRAM) is proposed for embedded applications. It is implemented using Si-TFETs and MIM capacitors using 28nm FDSOI-CMOS process which allows co-fabrication of CMOS and TFETs.

In section II, TFET devices used in this work are described; followed by uDRAM bitcell and its operation in section III. Conclusion is provided in section IV.

II. TUNNEL FETS

Our TFET device characteristics, including the advantages and drawbacks with respect to CMOS have been widely explained and published in [16-19]. In literature, the reverse-biased output characteristics are called ‘unidirectional’, due to the fact that the gate looses the control over the device for high reverse bias $V_{DS}$ condition, i.e. $V_{DS} < 0$ for nTFET. Such characteristics are shown in Figure 1 including the schematic representation of charge injection mechanism. Proposed cell in this work utilizes such reverse characteristics.

![Figure 1 Schematic representations of the charge injection mechanisms for a. the hump; b. the flat region and c. the thermionic region, and (d) reverse biased output characteristics of the TFET with VGS step: 0.25V, highlighting the three distinct regions.](image)

III. PROPOSED uDRAM

Proposed TFET 1T1C uDRAM cell is shown in Figure 2. In terms of architecture, it is similar to standard CMOS 1T1C DRAM cell. The cell is designed with the condition such that, $I_{OFF}$ (TFET off state current) $<< I_{LEAKCap}$ (Capacitor leakage current) $<< I_{HUMP}$ (TFET hump current due to NDR property of TFET). In such case, the TFET bitcell behaves like a static storage during retention with bitline (BL) at ‘0’, virtual ground node (G) connected to capacitance at 0.5V and wordline (WL) high at 1V. With this condition, while storing ‘0’ as shown in Figure 2a, access transistor ($T_A$) is in region I (refer Figure 1) with low $V_{DS}$ (0V) and high $V_{GS}$ (1V). In this condition, $I_{LEAKCap}$ will try to charge the node Q; however, node Q starts discharging due to $I_{HUMP}$ current through $T_A$ as soon as voltage on Q $> 0.5V$. Since $I_{HUMP} >> I_{LEAKCap}$, the cell maintains 0V, storing ‘0’ statically. In the case of storing ‘1’, the voltage across $C_s$ is ‘0’ and $T_A$ is in reverse bias with $V_{DS} = -0.5V$. Node Q will discharge due to $I_{OFF}$ current through $T_A$ while $I_{LEAKCap}$ will start charging node Q as soon as it goes below node G voltage (i.e. 0.5V). Since, $I_{OFF} << I_{LEAKCap}$, value of 0.5V on node Q is stored statically. Due to the static nature of storage during retention, no refresh is required which provides significant improvement in terms of throughput and energy consumption. Bitcell array organization is shown in Figure 3.

![Figure 2 TFET DRAM bitcell storing logical '0' and '1'](image)
B. Read Operation

Read operation is performed by pulling down the WLs for all the rows except the one selected for reading and pre-charging BLs to 0.5V. Figure 6 is showing the cells read and partially selected due to pre-charged BLs during read operation. The read operation waveform is shown in Figure 7 for reading ‘0’ and ‘1’. For reading ‘1’, BL remains on the pre-charged value of 0.5V while BL discharges for the bit having data ‘0’. It can be noted that read operation destroys the data ‘0’ in the cells. Therefore it has to be written back in the end of access, i.e. while closing the page as done in standard DRAMs. Signal voltages during read operation are shown in Table 2.

![Figure 6 Signal voltages during read operation](image)

Table 2 Signal voltages during read operation

<table>
<thead>
<tr>
<th>Operation</th>
<th>G</th>
<th>BL</th>
<th>WL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selected cells (i.e. one row)</td>
<td>Write ‘0’</td>
<td>Pulse</td>
<td>0V</td>
</tr>
<tr>
<td></td>
<td>Write ‘1’</td>
<td>(0.5V – 1.5V)</td>
<td>1.5V</td>
</tr>
<tr>
<td>Unselected cells</td>
<td>-</td>
<td>0.5V</td>
<td>1.5V or 0V</td>
</tr>
</tbody>
</table>

C. Bitcell Implementation

Bitcell is designed with planar access transistor and MIM capacitor. BLs are in metal 1 and WLs in poly with MIM capacitance. BL discharge occurs due to the charge sharing between the BL capacitance and node capacitance Cs. Unlike in CMOS DRAMs where Cs is decided by the leakage through the access transistor/capacitor and retention time requirement, our Cs requirement is relaxed and Cs equal to bitline capacitance is implemented in order to get 0.25V discharge on BL by charge sharing after WL activation. This allows to reduce the Cs significantly by 70%-85% and 40%-60% in comparison to conventional DRAMs and eDRAMs, respectively. During read/write operations the unselected rows of an array are having floating storage nodes because of access transistor (TA) in the bitcells is OFF with WL voltage of 0V. Once the access is finished, the 0V (logical ‘0’) is retained/restored by the TA, while, value of logical ‘1’, i.e. 0.5V, is always restored by ILEAKCap through Cs as previously explained for cell retention.
capacitor [2, 6, 7] above the devices. Unlike the high EOT of the capacitors in [2], which is used to reduce leakage, the proposed design uses EOT of 0.3nm which results in reduced area and higher leakage in the capacitor. With column size per BL of 128, extracted BL capacitance including wiring parasitics and device capacitances is ~2.5fF. The area of the bitcell is 0.0275μm² with logic design rules of 28nm FDSOI CMOS process. The estimated area with compact design rules for memories with planar transistor is 0.022 μm², showing an improvement of 8% to 10%.

D. Performances

Because of refresh removal, 3.8% and 7.8% throughput is gained in the best case of standard CMOS DRAM (i.e. assuming minimum penalty because of refresh) for less than 85°C and above 85°C, respectively. In an actual running system, the loss because of refresh cycles, which are interrupting running traffic on DDR, results in almost 8% to 18% for low and high temperature operation, respectively.

Sub-array dynamic power consumption during write is increased by 23% in comparison to standard CMOS DRAM, mainly due to the switching of virtual ground. However, overall the dynamic power consumption is lowered for the memory due to the energy gain because of refresh removal. Leakage current in the design is <1μA/bit on average, assuming 50% logical ‘1’ and logical ‘0’ storage in the memory, which is > 2 decades below in comparison to eDRAM [5] and up-to 48x lower in comparison to DRAMs [8] without using negative wordline and thick EOT for capacitances. Design summary and comparison with state-of-the-art is shown in Table 3.

It should be noted that the explanation in this paper uses particular voltages on signals as an example, these voltages can be tuned without any limitation to match the system requirements. The minimum requirement of the proposed design is to have three supply voltages, i.e. low, medium and high voltage supplies.

Table 3 Comparison

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Technology</td>
<td>28nm</td>
<td>22nm</td>
</tr>
<tr>
<td>Bitcell Area</td>
<td>0.0275 μm²</td>
<td>0.029μm²</td>
</tr>
<tr>
<td>Refresh Req.</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Retention Time</td>
<td>-</td>
<td>100μs</td>
</tr>
<tr>
<td>Capacitance</td>
<td>2.5 fF/cell</td>
<td>14.2 fF/cell</td>
</tr>
</tbody>
</table>

IV. CONCLUSION

A refresh free TFET bitcell is proposed utilizing negative differential resistance property of TFETs. Full memory architecture and its implementation using proposed bitcell is presented in this paper. Throughput gains of 12% to 18% can be achieved by using proposed architecture in Dual Date Rate (DDR) memories. Size of storage node capacitance (Cs) is reduced by 70%-85% and 40%-60% in comparison to conventional DRAMs and eDRAMs, respectively. Bitcell area of 0.0275μm² is achieved in 28nm FDSOI-CMOS using logic design rules which is scalable with technology shrink. The designed memory is compatible with DDR 1600 standard timings while removing refresh requirement. However, the proposed architecture is scalable and can be tuned in speed to meet requirements by either adjusting the sub-block sizes or voltages.

References