3T-TFET bitcell based TFET-CMOS Hybrid SRAM design for Ultra-Low Power Applications

Navneet Gupta[1,2], Adam Makosiej[2], Andrei Vladimirescu[1], Amara Amara[1], Costin Anghel[1]
[1] MINARC Laboratory, Institut Superieur 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— **This paper presents a TFET/CMOS hybrid SRAM architecture designed to address the requirements for ULP (Ultra-Low Power) applications, like IoT (Internet of Things). A novel 3-Transistor TFET SRAM cell is used for array while CMOS for periphery. The simulation extractions for power and speed are done including wiring and device parasitic capacitance from 4Kb SRAM designed in 28nm FDSOI CMOS process using MOSFETs & Tunnel FETs (TFETs). The proposed 3T-TFET SRAM cell supports aggressive voltage scaling without impacting data stability and allows application of performance boosting techniques without impacting cell leakage. A 0.35 fA/bit memory array leakage current was achieved showing a 14x to 10^4x improvement compared with state-of-the-art TFET and CMOS SRAM bitcells. Minimum read and write access pulse is evaluated at 1.27ns at sub-1V supply voltage.**

Keywords—Tunnel FET; Negative Differential Resistance(NDR); SRAM; Internet of Things(IoT)

I. INTRODUCTION

Constant increase in market demand for longer battery life and better performance particularly for devices in the IoT world forces the development of more energy efficient solutions. Since in modern SoCs SRAM can dominate the overall power consumption its power optimization is becoming ever more important [1-4]. In [4] Hanson et al. reported processors with 39% and 51% of total standby power dissipated in instruction memory and data memory, respectively, the SRAM being responsible for 90% of the total standby power consumption. In order to address this problem two main axes of improvement are being investigated, one focusing on architectural improvements and various circuit techniques and the other on introducing other than CMOS devices into SRAMs.

Owing to the increasing importance of the IoT market, many recent reports were oriented towards an optimization of standby power for systems with low activity durations [1-2]. To this end, efforts were made to optimize dynamic power consumption at both system and circuit level using various circuit techniques, like dynamic voltage frequency scaling (DVFS), power gating, etc. and/or sacrificing bitcell area to provide less leakage while maintaining sufficient performance [1-2]. DVFS is a particularly important technique in the IoT world due to the mode dependent operating frequency requirement, ranging from a few kHz to tens of MHz. In [1] the use of a ultra-long channel and ultra-low leakage bitcell is proposed, which together with an application of various circuit techniques allowed the reduction of SRAM leakage to only 27fA/bit. In [2], the authors proposed the use of a 10T-SRAM cell enabling low voltage operation and leakage reduction resulting in only a 30% of power consumption coming from the memory. Furthermore the separation of SRAM into two 4Kb macros allows application of independent active/retention operation modes providing system level control for SRAM power reduction. In [4], techniques like data compression and instruction set optimization were analyzed in order to reduce memory size requirement and standby power consumption.

Another approach to SRAM power optimization consists in exploring other than CMOS technology solutions for SRAM design. In [5], Baumann et al. used FeRAMs to reduce power consumption. However, FeRAMs need extra processing step for fabrication and additional control for operation resulting in extra cost and power budget. The Tunnel Field Effect Transistor (TFET) was proposed as a possible solution to reduce power dissipation. Tunnel transistors are different in their working principle from 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 [6-8]. Fabricated TFETs with S as low as 30mV/dec have already been measured [8]. 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. Saripalli et al. demonstrate in [9] a number of structures for heterojunction TFET SRAM simulated at VDD=0.3V. Yang et al. focused on the analysis of the 6T architecture with various kinds of transfer transistors (inward or outward n- or p-TFET), and on the analysis of the efficiency of assist techniques for VDD=0.8V [10]. Kim et al. demonstrated a novel 7T architecture simulated using heterojunction TFETs at VDD=0.5V [11], where the extra transistor serves as the read port, separating the read and write mechanisms, similarly to the CMOS 8T SRAM cells [12]. Yet another 6T cell structure operating at VDD=0.3V was demonstrated by Singh et al. [13], where the 6T cell core itself is modified to account for the particular operation of the TFETs.

Part of the above-mentioned reports on TFET SRAMs revealed difficulties in obtaining sufficient stability in read and write operations [9-11]. As the stability in both operation modes is inherently low due to the electrical performance of the TFETs, it is difficult for circuit designers to find the best balance between read and write. Moreover, due to the unidirectional TFET behavior, the researchers were forced to

target low-VDD operation resulting in even more difficulty in obtaining sufficient stability margins in active mode. New architectural solutions were developed [11, 13] to improve stability of single cell. But cells proposed in [11, 13] suffer from half-selection (HS) and write-disturb (WD) when organized as an array of memory cells.

Single and dual port TFET SRAMs proposed in [14,15] are free from HS and WD. In [15], TFET/CMOS hybrid Dual-Port SRAM (DPSRAM) based scratchpad memory is proposed with ultra-low leakage current (<5 fA/bit) with 29% area increase in comparison to a 6T-SRAM. However, dynamic power consumption for these memories is more than 6T-CMOS because of increased load on high capacitance nodes, like wordlines.

In [16], 4T-TFET SRAM cells is proposed using negative differential resistance (NDR) property of TFETs in reverse bias. The NDR property of TFET [17] is very promising in order to design compact latch. The architecture proposed in [16] suffers from stability and performance issues. In order to maintain data during read operation, read current should be less than the hump current (in pA range) provided by NDR. Such constrain leads to an extremely slow read with the risk of data corruption while executing the operation. Besides, the TFET transmission gate for data access limits the maximum operating voltage.

This work targets the investigation of ultra-compact SRAM design using Si TFETs, compatible with CMOS, for ULP with ultra-low leakage for long battery life time and good performance. We analyzed the architecture-level issues in TFET SRAM design and demonstrate a novel 3T-TFET SRAM cell designed using NDR property of TFETs in reverse bias. Proposed design supports aggressive voltage scaling without impacting data stability of the cell, and allows application of performance boosting techniques without impacting cell leakage. The new cell maintains reasonable stability in all operation modes without using any assist technique. With this, TFET/CMOS hybrid memory architecture is proposed using 3T-TFET bitcell with CMOS peripheral circuits. These two technologies are compatible for fabrication in a single FDSOI CMOS process. Our analysis remains valid for hetero-junction TFET designs as well.

II. TUNNEL FETS

TFETs are reverse-biased p-i-n gated junctions that operate by tunneling effect. Our devices were calibrated on data present in the literature [14, 17]. The structure is built using Low-k (SiO2) Spacers and a High-k (HfO2) Gate dielectric [17]. The gate and the spacers lengths are 30nm each. The gate dielectric physical thickness is 3nm, whereas the Silicon film (tSi) is 4nm. The TFET forward charactersitics, including the advantages and drawbacks with respect to CMOS have been widely explained in the literature [14-17]. The cell proposed here is based on typical reverse-biased output characteristics of TFET. Such characteristics are presented in Figure 1. Three regions can be distinguished in Figure 1 as follows: i. the hump; ii. the flat-current region and iii. the p-i-n turn-on.. The band-to-band tunneling current dominates the hump, the charge injection mechanism being schematically shown in Figure 2(a). The current severely reduces and attains its minimum in the flat-current region, whereas the tunneling current is suppressed due to the non-overlapping bands, as shown in Figure 2(b). The turn-on of the p-i-n diode is dominated by the thermionic emission over the barrier, a mechanism represented in Figure 2c.

In literature, the reverse-biased output characteristics are called 'unidirectional', due to the fact that the gate loses the control over the device for high negative drain voltages. Therefore, TFETs should not be biased in reverse with high negative V_{DS} for n-TFET to avoid high leakage currents, as shown in Figure 1. TFET gate to source (C_{GS}) and gate to drain (C_{GD}) capacitances are shown in Figure 3. The C_{GS} for TFETs is always low and has weak dependence on gate voltage; total gate capacitance is dominated by C_{GD}.

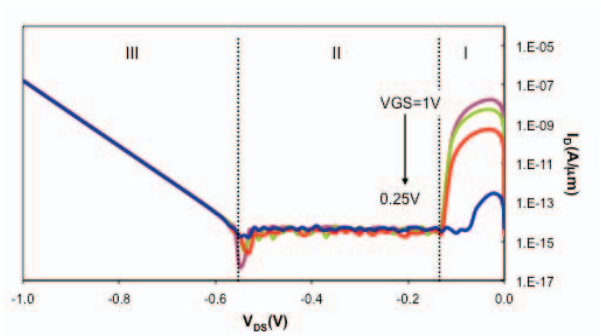
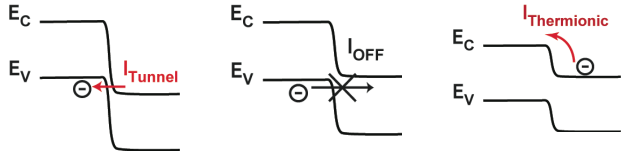


Figure 1 Reverse biased output characteristics of the TFET with V_{GS} step: 0.25V, highlighting the three distinct regions (see Section II).



a. b. c.

Figure 2 Schematic representations of the charge injection mechanisms for a. the hump; b. the flat region and c. the p-i-n turn-on region.

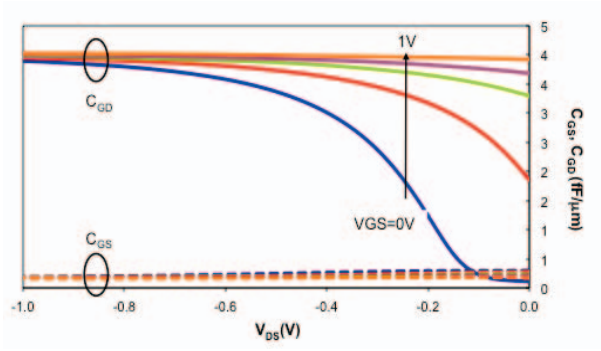


Figure 3 TFET Capacitances (C_{GS} and C_{DS}) for reverse bias V_{DS}

For circuit simulation, both p and n TFETs were modelled using look-up tables. Both DC and capacitance characteristics were implemented as $I_D(V_{GS},V_{DS})$, $C_{GS}(V_{GS},V_{DS})$, $C_{GD}(V_{GS},V_{DS})$ tables. From the analysis presented, the reduction in static power using TFETs is clear, however the limitation is in speed of operation and dynamic power consumption.

III. Proposed SRAM Cell

The basic idea of creating static latch using NDR with two TFET devices and SRAM bitcell using three TFETs is shown in Figure 4. During retention, the devices M_0 (PTFET) and M_1 (NTFET) are in reverse bias condition with $0 < V_D\text{-}V_S \leq 0.6V$ and BiasM0/BiasM1 are kept such that both devices get sufficient gate drive for hump. The I_D vs. Qint characteristics with reverse bias V_{DS} for two TFET devices connected in series is shown in Figure 5. This results in a latch behavior with the condition that the total cell supply should be less than the critical point where the TFET current becomes independent of gate voltage (refer Figure 1). For our devices this point is at 0.6V. Read operation is done using RBL and RWL lines with RBL pre-charged and RWL active low. Write operation is performed using a combination of voltages on supply and bias lines. Voltages of different signal during retention and write operations are shown in Table 1. Various operating modes are described in following sub-sections.

A. Retention Mode and stability

Figure 5 shows I_D for M_0 and M_1 with a sweep on Qint for circuit shown Figure 4(a). Since M_0 conducts near '0' and M_1 conducts near V_{DD}, '0' value in the cell is stored on M_0 with M_1 OFF, , '1' value in the cell is stored on M_1 with M_0 OFF. For cell supply of 0.6V, node Qint will be discharged through M_0 for 0 < Qint <100mV till Qint=0V. Similarly, Qint will be charged by M_1 for 0.5V < Qint < 0.6V. V_{Margin} shows the voltage range for which the cell is metastable and the distance between two stable states of the cell. The current peak value varies with the applied gate voltage but the width of the hump remains fairly independent of gate voltage (see Figure 1). The stability constraints for the proposed cell are significantly different from conventional 6T-SRAM cell because the data storing node (Qint) is isolated in all operating conditions. Therefore, stability during read/write operation is similar to static noise margin of the cell. This results in weak dependence of cell static noise margin on cell supply voltage. This cell has static noise margin of 100 mV (width of current hump) for $V_{Margin} \geq 0$.

Two variants of the proposed cell are shown Figure 4(a) and (b). Circuit shown in Figure 4(a) provides high write speed and low capacitance on supply nodes (V_D & V_S) because of low C_{GS} in TFETs and constant V_{GS} during write for M_0 and M_1. In circuit shown in Figure 4(b), V_{GS} for M_0 and M_1 is continuously changing with Qint during write operation because of source nodes connected to Qint. This results in performance penalty; however, this circuit is better in terms of stability because of higher hump current due to higher V_{GS} for M_0 and M_1 during retention mode.

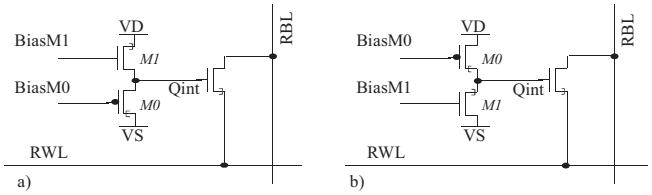


Figure 4 Proposed 3T-SRAM Cell Architectures

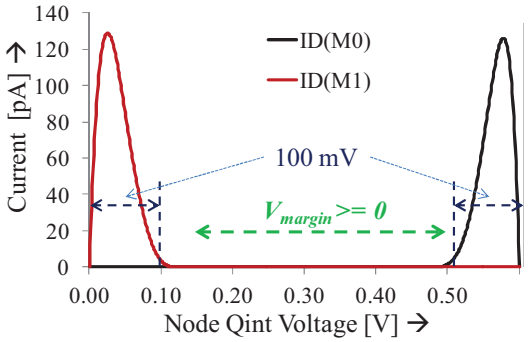


Figure 5 I_D for DC sweep on Node "Qint" at 0.6V Cell supply

Table 1 Supply voltages for different modes of operation

	Arch-1(a) and Arch-2(b) [V]				Arch-1 [V]		Arch-2 [V]	
	V_D	V_S	RWL	RBL	BiasM0	BiasM1	BiasM0	BiasM1
Retention	0.75	0.25	0.75	0.75	0	1	0	1
Write-1	0.25	0.75	0.75	0.75	0	0.5	0.5	1
Write-0	0.25	0.75	0.75	0.75	0.5	1	0	0.5
Read	0.75	0.25	0	0.75	0	1	0	1

B. Write Operation:

During retention, M_0 and M_1 are in reverse bias. Basic idea for writing in this cell is to forward bias both M_0 and M_1 by changing V_{DS}, such that they behave like FETs, and control the gate bias to switch off either M_0 or M_1 depending on the value to be written in the cell. Different signal voltages during retention and write are shown in Figure 6. For writing in the cell shown in Figure 4(a) (Arch-1), V_D and V_S voltages are swapped to make $V_S > V_D$ and both M_0 and M_1 are forward biased. For writing '0', BiasM0 is pulled up to reduce the gate drive of M_0 and BiasM1 remains the same; thus, M_1 will discharge the node Qint to voltage on V_D. Similarly for writing '1', BiasM1 is pulled down to reduce gate drive of M1, BiasM0 remains same; therefore, M0 will charge the node Qint to the voltage on V_S. Waveforms for writing '0' and '1' are shown in Figure 7.

C. Read Operation

Read operation is done using single ended read scheme with RWL and RBL lines. RWL selects the row to be read. RBL can be allowed to discharge fully or a single-ended sense amplifier can be used. Full discharge is preferable for low voltage operation and to maintain the column pitch for reading circuit. The read waveform is shown in Figure 8 for reading '0' and reading '1'.

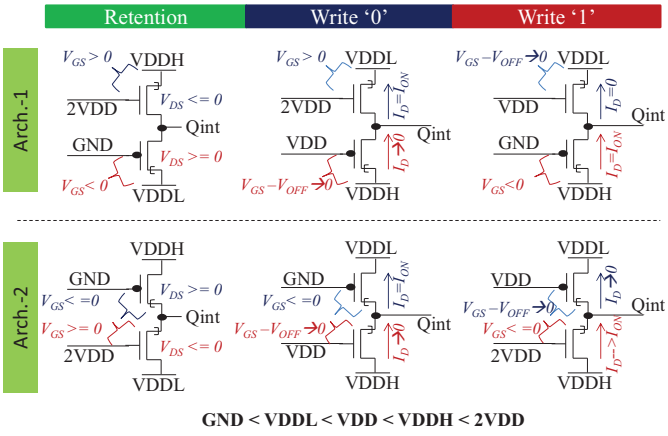


Figure 6 Signal Voltages for different cell states

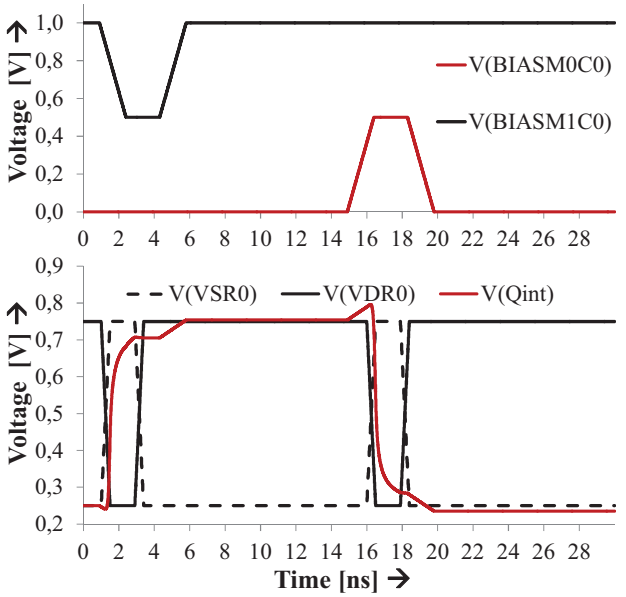


Figure 7 Write Operation (Write-'0', Write-'1')

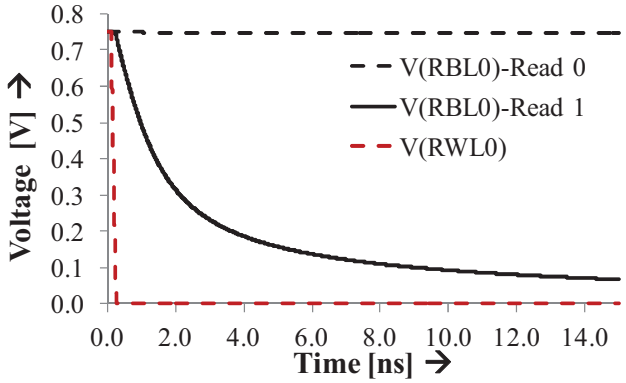


Figure 8 Waveform-Read operation

IV. Memory Architecture

Figure 9 shows the memory cell array organization including routing of signals. Data words are stored horizontally. V_D and V_S are routed horizontally to align with the data word. BiasM0 and BiasM1 are routed vertically. In this architecture, selection of the row to be written is done by V_D and V_S, and value to be written in each cell is decided by BiasM0 and BiasM1. Selection of row to be read is done by RWL and data is read using RBL's.

Proposed memory architecture is shown in Figure 10. In order to optimize the cell array leakage current, bitcell array is designed fully with TFETs. Periphery is designed using CMOS to optimize area for the same speed of operation in comparison to TFETs because of higher drive strength. We have used single ended sense amplifier to limit the bitline discharge to reduce power consumption and to allow bigger column size.

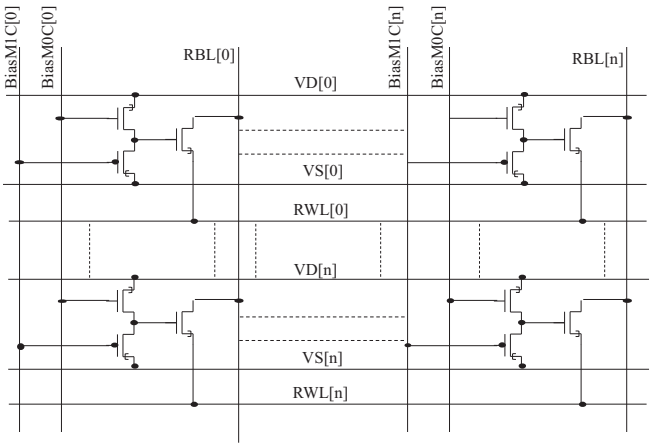


Figure 9 Bitcell and corresponding signal organization in SRAM cell array

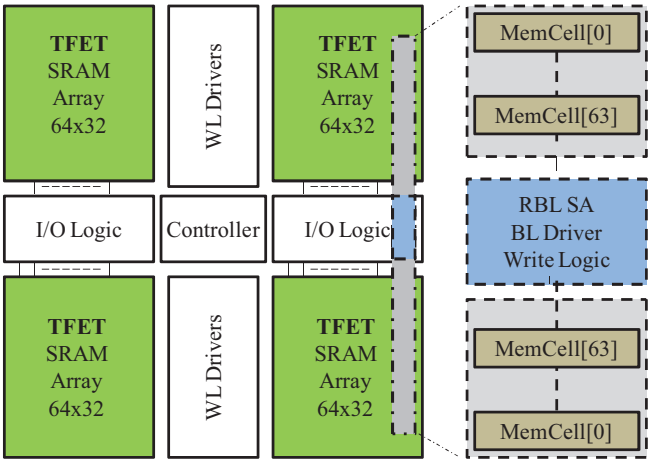
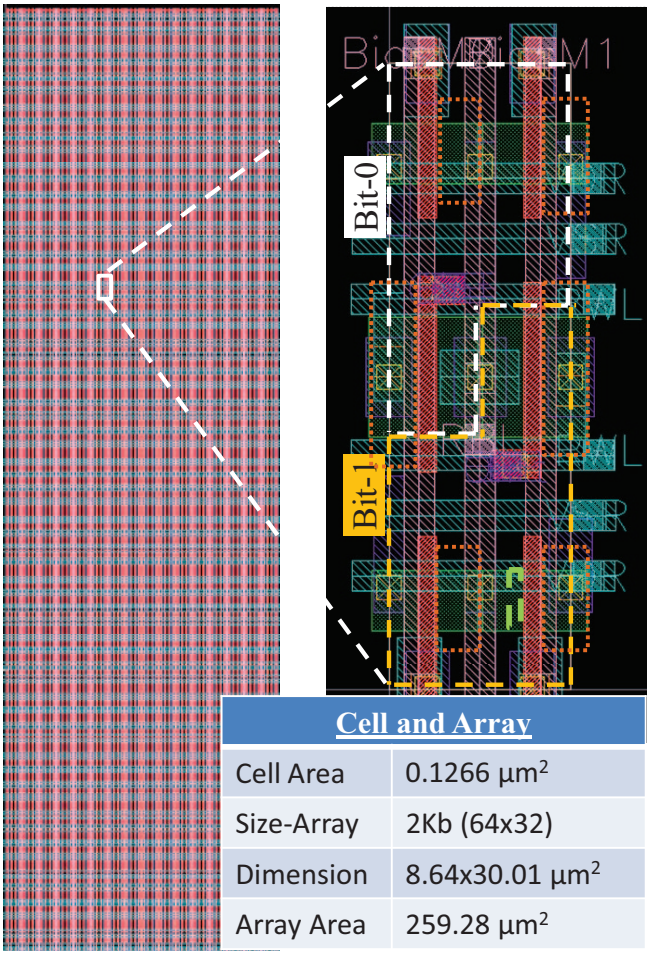


Figure 10 Proposed TFET/CMOS Hybrid Memory Architecture

A. Memory Layout

Layout for dual-cell block and cell array of 64x32 size is shown in Figure 11. The cell size is 0.1266 μm^2/bit with logic design rules. Area is similar to industrial high density (HD) 6T-CMOS cell which used compact design rules in comparison to logic design. In this layout, the TFET M_3 (200nm) on read port is twice the size of M_0 and M_1 (100nm). This improves the read speed of the design. In order to have optimized rectangular layout of bitcell, two bitcells are combined together in layout to have six transistors. Cell boundaries are represented by dotted lines to show each cell separately. Because of the reduced cell width, the wiring capacitances on the various horizontal lines in the cell array are reduced. The extracted values of wiring capacitances form the layout are 50% in comparison to the same size of memory designed using compact 6T-SRAM cell. Due to the low bit-cell capaciatnce and low C_{GS} capacitance of TFET devices, the total capacitance on V_D, V_S and RWL lines are less than half with respect to standard 6T-CMOS. This results in drivers (for RWL, VD and VS) with less leakage for same specification of transition time and word size.



Cell and Array	
Cell Area	0.1266 μm^2
Size-Array	2Kb (64x32)
Dimension	8.64x30.01 μm^2
Array Area	259.28 μm^2

Figure 11 Cell array and cell layout

B. Energy Efficiency

For the proposed design energy consumption is computed with the following assumptions, during read 50% of data are '0 and 50% are '1'; and 50% operations are read and 50% are write. Overall comparison of energy consumption in various modes and bitcell area is shown in Table 2. E_{READ} is the energy consumed during read on row drivers and bitlines. Bitline discharge is limited to 200 mV for 3T-TFET and 8T-TFET SRAMs because they use single ended sensing. Bitline discharge is limited to 100mV for 6T-SRAM with differential read. E_{WRITE} is energy consumed during write operation. I_{LEAK} in active mode is total leakage in bitcell array and periphery with dynamic power gating implementation (only 25% of the drivers are switched ON depending on the accessed address). I_{LEAK} in standby mode is computed with periphery OFF and cell array power ON to retain the data.

We have analyzed and compared the leakage power consumption of wordline drivers for proposed design with 6T-SRAM and 8T-TFET SRAM cells for same transition time specification. Bitcell array leakage is 10^4x and 77x lesser in comparison to 6T-CMOS (HD) and ultra-low leakage 6T-65nm CMOS SRAM [1] cells, respectively. Bitcell leakage is 14x lesser in comparison to 8T-TFET SRAM cells [14, 15]. During standby, total leakage is coming from cell array, thus TFET memory leakage is much lower than CMOS memories. Overall memory leakage during active mode, including bitcells and drivers, for proposed design is 52% less than 6T-CMOS and 77% less than 8T-TFET SRAMs.

Because of low capacitance on V_D, V_S and RWL lines, dynamic power consumption is our design is 70% less in comparison to 6T-CMOS SRAM and up-to 90% less in comparison to 8T-TFET SRAM.

C. Read and Write Performance

Read and write minimum wordline pulse width (WLP_{crit}) is shown in Figure 12. For the full range of operation cell leakage is < 0.35fA/bit because devices M_0 and M_1 are in reverse bias. As shown in Figure 12, the read/write speed can be increased by using assist techniques for low voltage operation. Since the cell data storage node is isolated from RBL and bias voltages of M_0 and M_1, negative wordline (NWL) can be used to increase read speed without impacting the cell stability. Similarly for write, speed can be increased without degrading cell stability by boosting bias voltages (BiasM0 and BiasM1). Read performance is improved by 29x, by using negative wordline (RWL) with -100mV and -150mV for 0.4V and 0.3V supplies, respectively. Similar for write, with a boost of 100 mV on bias voltage results in 4.8x improvement at 0.3V cell supply.

Overall performance is estimated including periphery delays in row decoder, drivers and sensing. Proposed design supports overall read speed from 1.92 GHz to 3.82 MHz and write speed from 429 MHz to 17.3 MHz for 0.6 V to 0.3V on cell supply, with BiasM0 and BiasM1 from 1.2V to 0.6V. This includes, overall five voltages. This can be implemented either with five supplies or three supplies with two voltage dividers.

Table 2 Comparison-Power and Area

Cell	VDD [V]	WL_{Pulse} Read[ns]	WL_{Pulse} Write[ns]	E_{READ} [fJ/acc.]	E_{WRITE} [fJ/acc.]	E_{AVG} [fJ/acc.]	I_{LEAK} Active [pA/bit]	I_{LEAK} STBY [fA/bit]	Area-Bitcell [μm^2]
8T-TFET[14]	1	0.26	1.10	28.5	61.0	44.8	25.5	5.00	0.336
6T-CMOS	0.75	0.27	0.16	9.10	12.9	11.0	11.8	$7.8*10^3$	0.120
3T-TFET[Proposed]	0.6	0.21	0.93	1.81	4.99	3.40	5.72	0.35	0.1266
6T-CMOS[1]	1.2	7ns (access time)*		N.A.	N.A.	25µW/MHz*	N.A.	27.0	2.04

*** Measurement values reported for full memory [1]**

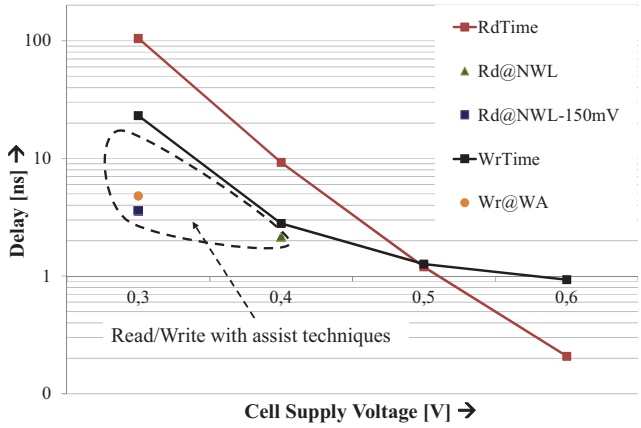


Figure 12 Read/Write performance (WLP_{crit}) vs. cell supply voltage, including improvements with read/write assist techniques

V. Conclusion

A new integrated 3T-TFET bitcell based TFET/CMOS hybrid SRAM architecture has been introduced. The memory array is built with TFETs for ultra-low leakage and the CMOS periphery. The proposed 3T-TFET bitcell uses NDR based 2T-TFET latch for data storage and 1T for read port. A new write mechanism using supply lines has been proposed. Ultra-low leakage current (< 0.35fA/bit) of memory cells has been achieved. Proposed design supports voltage scaling and works from 0.6V to 0.3V bitcell supply voltages. Read/write speed improvement up-to 70% and 90% can be done using assist techniques for 0.4V and 0.3V cell supply, respectively.

References

[1] Toshikazu Fukuda, Koji Kohara, et al.,"A 7ns-Access-Time 25µW/MHz 128kb SRAM for Low-Power Fast Wake-Up MCU in 65nm CMOS with 27fA/b Retention Current", ISSCC 2014

[2] Myers, James, et al. "An 80nW retention 11.7 pJ/cycle active subthreshold ARM Cortex-M0+ subsystem in 65nm CMOS for WSN applications." Solid-State Circuits Conference-(ISSCC), 2015 IEEE International. IEEE, 2015.

[3] Lim, Wootaek, et al. "Batteryless Sub-nW Cortex-M0+ processor with dynamic leakage-suppression logic." Solid-State Circuits Conference-(ISSCC), 2015 IEEE International. IEEE, 2015.

[4] Hanson, Scott, et al. "A low-voltage processor for sensing applications with picowatt standby mode." Solid-State Circuits, IEEE Journal of 44.4 (2009): 1145-1155.

[5] Michael Zwerg, et al., "An 82µA/MHz Microcontroller with Embedded FeRAM for Energy-Harvesting Applications", ISSCC 2011

[6] J. Appenzeller, et al., "Band-to-Band Tunneling in Carbon Nanotube Field-Effect Transistors", Physical Review Letters, Vol. 93, no. 19, pp. 196805-1, 2004

[7] Villalon, A., et al. "First demonstration of strained SiGe nanowires TFETs with ION beyond 700µA/µm." VLSI Technology (VLSI-Technology): Digest of Technical Papers, 2014 Symposium on. IEEE, 2014.

[8] R. Gandhi, et al., "Vertical Si- Nanowire n- type vertical tunneling FETs with low subthreshold swing (≤50 mV/decade) at room temperature", IEEE Electron Device Letters, Vol. 32, pp. 437-439, 2011

[9] V. Saripalli, et al., "Variation-Tolerant Ultra Low- Power Heterojunction Tunnel FET SRAM Design", IEEE/ACM International Symposium on Nanoscale Architectures (NANOARCH) 2011

[10] X. Yang and K. Mohanram, "Robust 6T Si tunneling transistor SRAM design", DATE 2011

[11] D. Kim, et al., "Low Power Circuit Design Based on Heterojunction Tunneling Transistors (HETTs)", ISLPED 2009

[12] L. Chang, et al., "Stable SRAM Cell Design for the 32 nm Node and Beyond", Symposium on VLSI Technology Digest of Technical Papers 2005

[13] J. Singh, et al., "A Novel Si-Tunnel FET based SRAM Design for Ultra Low-Power 0.3V VDD Applications", Design Automation Conference 2010

[14] A. Makosiej et al., "A 32 nm Tunnel FET SRAM for Ultra Low Leakage," ISCAS, 2012.

[15] N. Gupta et al., "Ultra-Low Leakage sub-32nm TFET/CMOS Hybrid 32kb Pseudo Dual-Port Scratchpad with GHz Speed for Embedded Applications," ISCAS, 2015.

[16] V. Saripalli etal., "Generic TFET based 4T memory devices," US Patent-2014, No. US8638591

[17] C. Anghel, et al., "30-nm Tunnel FET with improved performance and reduce ambipolar current", IEEE Transactions on Electron Devices, Vol. 58, pp. 1649-1654, 2011.