An ultra-low power 3-terminal memory device with write capability in the off-state

Xiaoyao Song, Ashwani Kumar, and Maria Merlyne De Souza[1]

[1] EEE Department, University of Sheffield, Sheffield, UK,
*email: m.desouza@sheffield.ac.uk

Abstract

In this work, we demonstrated a room temperature fabricated ZnO/Ta_2O_5 transistor for low power compute-in-memory application. By writing during the off-state, the device programmed for compute-in-memory shows power consumption in nW. By using variable pulse amplitudes for SET/RESET allows control of the on/off ratio of resistance states without affecting power consumption. Benchmarked against other ReRAMs the device shows a competitive 8 nJ per transition, which allows a reduction of power consumption in comparison to a filamentary device. (Keywords: Tantalum oxide, zinc oxide, oxygen vacancies, memory TFTs and compute-in-memory)

Introduction

Extremely low current consumption, switching speed of 10-100 nanoseconds and compatibility with CMOS have made ReRAM the default technology of choice for compute-in-memory and neuromorphic applications. However, ReRAMs rely on filamentary processes that face challenges of variability at extreme scaling and require write voltages (V_{write}) that are difficult to reliably scale at < 0.3V. As the on-state of the device typically has low resistance (~ 100 Ω), this results in a high current I_{write} $(\sim\mu A - mA)$ and high power consumption in $\mu W - mW$, although exceptions are reported in [1]. Moreover, since the ReRAMs are a type of two terminal device, the reading and the writing of the device cannot be performed simultaneously. This motivates the search for new physical mechanisms to address the gap between current computing performance and the ultimate efficiency of the human brain [2].

This paper presents a radically different concept in memory for SET/RESET, while the device is off. This is made possible by decoupling the programming within the gate from the conduction path in the channel of a 3 terminal synaptic device, allowing the device to perform reading and writing simultaneously. Our channel acts merely as a sensor of the state written in the gate insulator. We demonstrate compute-in-memory and estimate the scaling potential and applications of this concept.

Methodology

A. Device Fabrication and characterization

Our device is a Solid Electrolyte FET (SE-FET) $(W \times L = 100\,\mu m \times 1.5\,\mu m)$ fabricated via RF sputtering at room temperature and fully compatible with BEOL processing [3]. A conducting Indium Tin Oxide (ITO, 20 Ω/square) is used as gate metal. 275 and 350 nm of tantalum oxide (Ta_2O_5) with $\varepsilon \approx 20.8$ acts as bottom gatethe insulator, over which $40\,nm$ of ZnO is sputtered as channel. $100 \times a\,a\,a\,200\,\mu m^2$ Al contacts are thermally evaporated on ZnO. Electrical characterization is undertaken using Keysight B2902A.

Fig. 1 (a) highlights the principle of operation of our SE-FET in comparison to a conventional Resistive Switch (RS), consisting of 2-terminals separated by $\sim 10nm$ thick oxygen vacancy rich dielectric material. The resistance state of the ReRAM relies on filamentary processes, as shown in Fig. 1 (a). The writing stage of the device is usually done via application of a high voltage on the device to drive the vacancies to form the filament (or to eliminate the existing one) between the cathode and anode. During the reading stage, a relatively small voltage (V_{read}) is applied across its two terminals to check the current state of the device. The reading voltages tend to be small(<0.1V) and non-volatile flip-flops can help scale V_{read} to V_T or even sub-V_T levels [4]. However, the write energy is usually technology-dependent and more challenging to scale due to variability and reliability with the size of the conductive filament [5].

Fig. 1 (b) highlights our new concept of the SE-FET, where read and write occur across two separate terminals, Drain-Source and Gate-Source, respectively, of the transistor. The O_2 vacancies rich Ta_2O_5 layer is deposited using room temperature RF sputtering. Once a positive V_{GS} is applied on the gate, the negatively charged O_2 vacancies accumulate at the Ta_2O_5/ZnO interface attracting electrons in the ZnO channel, leading to an improved channel conductivity. This change in channel conductivity tends to be maintained in a similar way as a ReRAM. During the writing stage, the drain and source of the device are grounded, the device functions as a two terminal device where the current flows across the gate oxide. Due to the thick gate dielectric, despite a large V_{GS} required during writing, I_{write} remains limited by the low gate leakage current I_{GS} in nA, thereby facilitating orders of magnitude reduction in power consumption. During the reading stage, the drain source of the device functions as in a ReRAM whereby a small V_{DS} is applied across the channel to read the resistance state of the device.

Results

The transfer characteristics under dynamic gate bias sweep in Fig. 2 (a) show a memory window (counter-clockwise hysteresis), whose width becomes wider for thicker gate oxide. Moreover, the device with gate oxide thickness of $275\,nm$ exhibits a sub-60 mV/dec switching during the backward sweep, arising from an abrupt depletion of the accumulated oxygen ions during the reverse sweep [6]. Fig. 2 (c) shows measured I_{DS} arising from a series of bias pulses V_{GS} in Fig. 2 (b) at $V_{DS} = 0.1\,V$. Red circles in Fig. 2 (c) indicate points in time when $V_{GS} = 0$. The polarity of V_{GS} defines the low/high resistance states, (LRS/HRS), corresponding to accumulation/depletion of O_2 vacancies at the interface, as illustrated in the insets in Fig. 2 (b). The device is operated between two terminals V_G and $V_B (= V_D = V_S)$ during write and $V_{DS} = 0.1\ V$ with $V_{GS} = 0$ during read, (Fig. 2 (d)), as implemented via alternating pulses shown in Fig. 3 (a). Fig. 3 (b) shows the resultant measured I_{DS} for $T_{ox} = 275\ nm$, L=1.5 μm during read, revealing distinct LRS and HRS states, separated by at least 3 orders of magnitude. A higher resistance of $\sim 1\ M\Omega$ in LRS keeps the drain current low, thereby reducing the power consumption during the read stage. The inset shows the maximum and minimum of each state in these unoptimized devices that vary by less than an order in magnitude. Moreover, the standard deviation of HRS (0.3 of an order) remains higher than for LRS (0.06 order), as expected, due to the higher impact of noise on the HRS. Fig. 3 (c) shows the gate current characteristics vs. time, whose magnitude remains limited to $10\ nA$, establishing that the maximum power consumption during write is less than $50\ nW$ for $V_{GS} = 5\ V$.

B. Logic operations

Similar to a bipolar or complimentary resistance switch (BRS or CRS) [7], an example of the SE-FET as a finite-state-machine is depicted in Fig.2 (d). Application of V_G or V_B =1 or 0 shifts the device from LRS to HRS and vice-versa, whereas $S = 0$ or 1 represents whether the device is in HRS or LRS. Fig.3 (d) shows the truth table corresponding to the state diagram, where the previous state of the device serves as an additional input for determining the following state. From the truth table, the equation of state is

$$S = S \cdot V_G + S \cdot V_B' + V_G \cdot V_B' \qquad (1)$$

Where the operators $\cdot$, $+$, and $'$ refer to AND, OR, and NOT logic, respectively. 14 out of a possible 16 logical functions of two inputs can be performed on a single device, while the remaining two functions, namely XOR and XNOR require an additional device [7]. A single device can be used to perform logic involving two or multiple inputs provided their operation can be broken down into a series of steps, in accordance with Eq. (1). Fig.4 (a) highlights an example of the two steps required to realise material implication or IMP operation (if A then B), $A' + B$, which is an important operation to complete a logic set when combined with RESET [8]. In Step 1, S=1, while in Step 2, inputs A and B are applied to V_B and V_G with all possible outcomes summarized in the truth table. Fig.4 (b) shows the sequence of V_{GS} pulses generated as per the inputs V_G and V_B indicated in each row of the truth table in Fig.4 (a). The measured I_{DS} are plotted in Fig.4 (c), for a device with $T_{ox} = 275\ nm$, showing close agreement with the expected state in the last column in the truth table in Fig.4 (a) (shown by colour).

C. Benchmark

Table.1 summarises the characteristics of state-of-the-art devices in comparison with our experimental results measured at $250\ ms$ of switching time with $V_{GS} = 5V$. MTJs demonstrate fastest swithe ching, but OxRAM also switches in ns with energy scaling in nJ. Summarising the results of Table 1 alongside reported non-filamentary ReRAMs, there is evidence of a programming voltage-switching speed-current dilemma, especially with non-filamentary ReRAMs operating in the time scale of seconds or even hours [9] as current levels are reduced in the quest towards truly neuromorphic systems [10]. Other non-filamentary devices scaled x1000 reported in [11] for example have power consumption in μW range. In comparison, the current device operates with a common gate that extends over the entire chip and suffers due to 100% overlap capacitance with the source/drain pads and underlying gate oxide. Moreover, channel lengths are 1.5 μm and above. Plotting our experimentally obtained switching time versus channel length in Fig.5 shows the lowest power consumption/unit area reported to date. However, the switching speed is poor. Hence, extrapolating to similar dimensions as contemporary ReRAMs based on measured data as a function of channel length, in Fig. 6, shows the potential of a switching speed of 150 ns, well within reach of a vertical cross-bar array using nanowires for this device.

Potential applications of the SEFET include spike sorting in neural (or brain-computer) interfaces, where limited communication bandwidth and energy budget prohibit the transmission of large volumes of neuronal data for sorting in real time. Our potentially flexible platform technology is ideally suited for hardware implementable spike recording and feature extraction, as it does not even require historical data storage, at speeds of a few 100 kHz. More recently, even volatile memory has been shown to serve this purpose [12].

Conclusion

A new concept for reducing power consumption is demonstrated in an ionic device with a non-filamentary mechanism by writing the device in the off-state. At least 3 orders of magnitude difference in the HRS and LRS states, with power consumption in nW despite large operating voltage is demonstrated with experimental examples of logic operation.

Acknowledgements

The devices were fabricated by Dr. P. B. Pillai. MMS acknowledges partial support from ENIAC project E2SG.

References

[1] "Stanford Memory Trends." https://nano.stanford.edu/stanford-memory-trends. [2] E. J. Fuller *et al.*, Adv. Mater., 29 (4), 1604310, 2017. [3] P. Balakrishna Pillai *et al.*, ACS Appl. Mater. Interfaces, 9(2) 1609, 2017. [4] I. Kazi *et al.*, IEEE Trans. Circuits Syst. I Regul. Pap., 61(11), 3155, 2014. [5] D. Ielmini, Semicond. Sci. Technol., 31(6), 1, 2016. [6] A. Kumar *et al.*, ACS Appl. Mater. Interfaces, 10(23), 19812, 2018. [7] E. Linn *et al.*, Nanotechnology, 23(30), 305205, 2012. [8] J. Borghetti *et al.*, Nature, 464(7290), 873, 2010. [9] I. Sugiyama *et al.*, *APL Mater.*, 5(4), 046105, 2017. [10] J. Zhou *et al.*, *IEEE Electron Device Lett.*, 37(4), 404, 2016. [11] B. Govoreanu *et al.*, *2016 IEEE Symposium on VLSI Technology* 2016. [12] I. Gupta *et al.*, *IEEE Trans. Biomed. Circuits Syst.*, 12(2) 351, 2018. [13] T. Breuer *et al.*, *Adv. Electron. Mater.*, 1(10), 1500138, 2015. [14] T. Breuer *et al.*, *Sci. Rep.*, 6, 23967, 2016. [15] P. Huang *et al.*, *Adv. Mater.*, 28(44), 9758, 2016. [16] A. Jaiswal *et al.*, *Sci. Rep.*, 8(1), 1, 2018. [17] M. Jin, L *et al.*, *Nanotechnology*, 29(38), 385203, 2018.

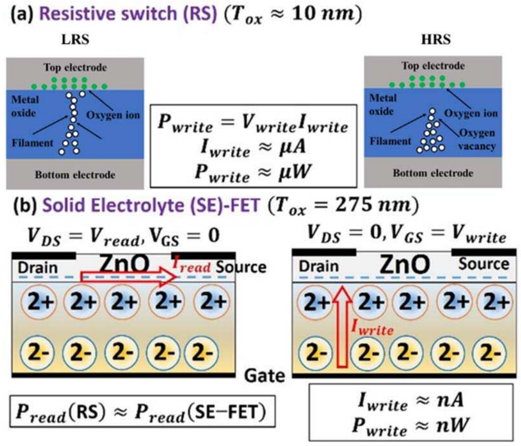

(a) Resistive switch (RS) ($T_{ox} \approx 10\ nm$)
LRS
HRS
Top electrode
Metal oxide
Oxygen ion
Filament
Bottom electrode
Oxygen vacancy
$P_{write} = V_{write} I_{write}$
$I_{write} \approx \mu A$
$P_{write} \approx \mu W$
(b) Solid Electrolyte (SE)-FET ($T_{ox} = 275\ nm$)
$V_{DS} = V_{read}, V_{GS} = 0$
$V_{DS} = 0, V_{GS} = V_{write}$
Drain
ZnO
I_{read}
Source
Gate
I_{write}
2+
2-
P_{read}(RS) ≈ P_{read}(SE−FET)
$I_{write} \approx nA$
$P_{write} \approx nW$

Fig. 1. Mechanism of read and write operations in (a) resistive switching (RS) and (b) the solid-electrolyte field effect transistor (SE-FET), demonstrating the possibility of low power consumption during writing.

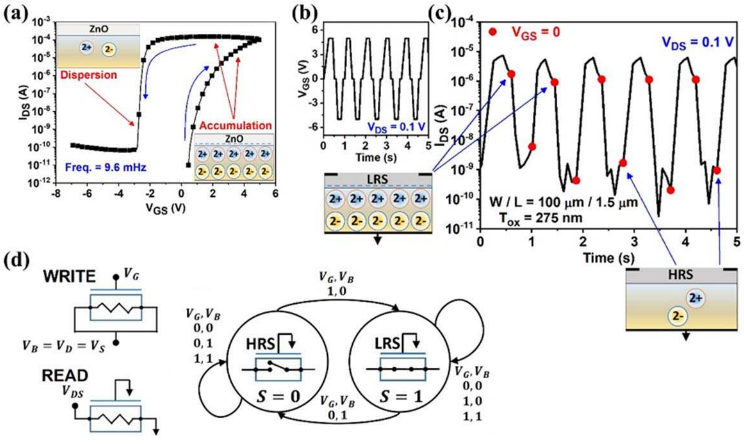
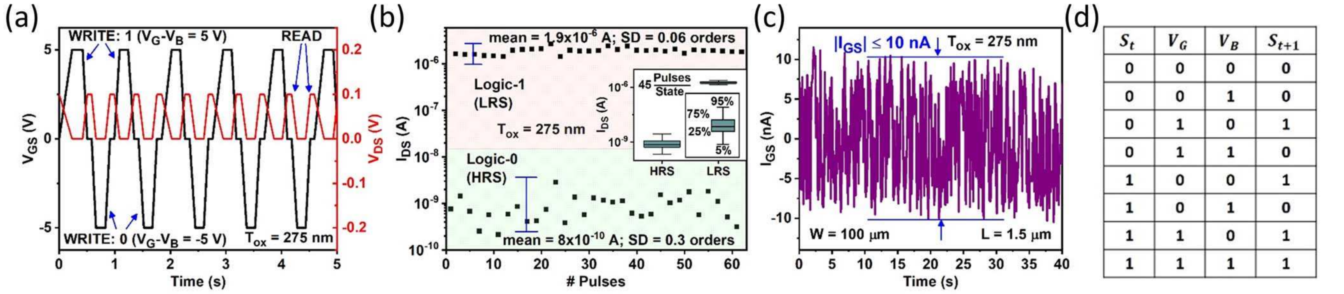
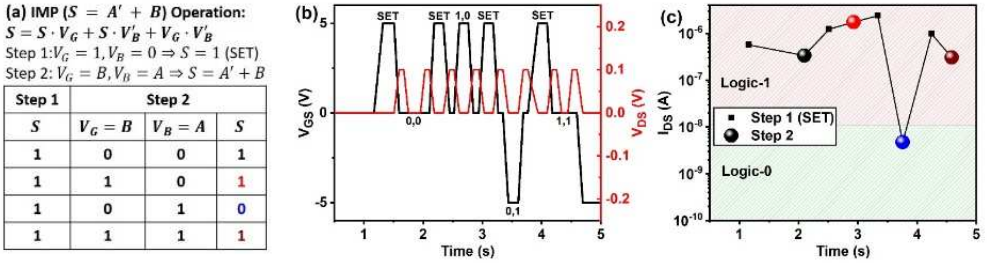


Fig. 2. Measured (a) $I_{DS} - V_{GS}$ under dynamic gate bias sweep of devices at two oxide thicknesses, (b) gate bias pulses applied to the device with a constant $V_{DS} = 0.1\ V$, (c) drain current characteristics as a function of time, exhibiting two resistance states (HRS and LRS) as gate bias crosses zero, indicated by red circles. (d) The bias configurations and finite state diagram of the device, where a 1 for either V_G or V_B indicates an application of a positive bias, S=0 and 1 indicate the LRS and HRS.



S_t	V_G	V_B	S_{t+1}
0	0	0	0
0	0	1	0
0	1	0	1
0	1	1	0
1	0	0	1
1	0	1	0
1	1	0	1
1	1	1	1

Fig. 3. (a) The scheme of bias pulses for V_{GS} and V_{DS}, such that V_{DS} remains zero during write operation ($V_{GS} \neq 0$) and vice versa, ensuring "write" and "read" operations occur exclusively to each other. (b) The resulting measured drain current during "read", indicating the presence of two resistance states (HRS and LRS) separated by three orders of magnitudes, with small variance (inset). (c) Measured gate current characteristics, establishing that power consumption remains limited to nW during "write", since no drain current is zero. (d) The truth table for describing the state transition.


(a) IMP ($S = A' + B$) Operation:

$S = S \cdot V_G + S \cdot V'_B + V_G \cdot V'_B$

Step 1:$V_G = 1, V_B = 0 \Rightarrow S = 1$ (SET)

Step 2: $V_G = B, V_B = A \Rightarrow S = A' + B$

Step 1	Step 2		
S	$V_G = B$	$V_B = A$	S
1	0	0	1
1	1	0	1
1	0	1	0
1	1	1	1

Fig. 4. (a) Steps to perform the material implication (IMP) (if A then B) logic operation and corresponding truth table, (b) applied gate and drain bias pulses and (c) corresponding measured drain current for each of the steps indicated in (a). W×L=100×1.5 μm^2, Tox=275 μm.

Table 1. Comparison of the Ta2O5/ZnO based SE-FET with state-of-the-art two-terminals devices reported to date. Despite a larger pulse-width, the measured power consumption during the transition is of the order of nW, owing to writing without turning the device ON in the SE-FET.

Description [Reference]	R_{LRS}	R_{HRS}/R_{LRS} (10^x)	Switching Time	Pulse Amp. (V)	Energy Consumption	Approx. Power in a transition
HfO_2 complimentary RS (CRS) [11]	10-100 $k\Omega$	3	0.5μs	1	0.5 nJ	1 mW
Ta_2O_5 CRS [12]	4 $k\Omega$	2-3	3 ms	2.4	3 μJ	0.6 mW
HfO_X/AlO_Y RS [13]	1 $k\Omega$	2	100 ns	1.2/-1.8	0.324 nJ	3.24 mW
Magnetic Tunnel Junction [14]	6.4 $k\Omega$	0.5	2 ns	0.8	0.067 pJ	33.5 μW
GeTe RS [15]	10 $k\Omega$	3-4	60 ns	1.2	498 pJ	8.3 mW
amorphous In-Zn-Sn-O RS [16]	12 $k\Omega$	2	100 ns	3	75 pJ	0.75 mW
Al_2O_3/aSi RS [9]	3.6 $G\Omega$	2	10 ms	3.6	36 pJ	3.6 nW
Nanohole graphene/HfO_2 RS[17]	1 $k\Omega$	6	100 ns	4	1.6 nJ	16 mW
Ta_2O_5/ZnO SE-FET [This Work]	**1 $M\Omega$**	**3**	**250 ms**	**5**	**8.1 nJ**	**32.4 nW**

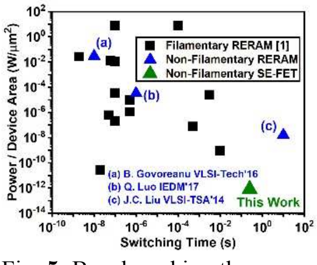


Fig. 5. Benchmarking the power consumption per unit device area vs. switching time from the experimental data reported in [1] and our current measurements.

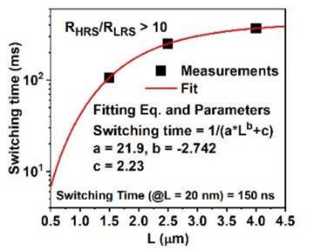


Fig. 6. Measured switching time vs. gate length L. The fit to data predicts 150 ns of switching time for $L = 20\ nm$.