Extended Abstracts of the 2018 International Conference on Solid State Devices and Materials, Tokyo, 2018, pp23-24

Time-resolved conductance in electrochemical systems for neuromorphic computing

Douglas Bishop, Paul Solomon, Seyoung Kim, Jianshi Tang, Jerry Tersoff, Teodor Todorov, Matt Copel, John Collins, Ko-Tao Lee, Sanghoon Shin, Wilfried Haensch, John Rozen

IBM Research, TJ Watson Research Center, 1101 Kitchawan Rd, Yorktown Heights, NY, 10598, USA
Phone: +1-914-945-3148, E-mail: dmbishop@us.ibm.com

Abstract

While Electro-Chemical Random Access Memory (ECRAM) devices demonstrate promising non-volatile analog symmetry compared to RRAM & PCM synaptic elements, their response at short time scales is key for their usefulness in neuromorphic computing. For the first time, we measure the time-resolved dynamic behavior at sub-μs resolution and show successful programming with 100 ns write pulses. Observed device transients are on the μs and ms timescale for larger devices. Equivalent circuit model and physical origin of the transients are proposed. We show how scaling, engineering and material changes can be used to reduce the transients and provide a viable element for use as a synaptic switch, at target speed, in analog neuromorphic computation.

1. Introduction

Three-terminal ECRAM has demonstrated deterministic multi-level symmetric conductance changes upon ion intercalation in the channel, high on/off ratios, and low power characteristics [1], which are key weight update properties for deep learning acceleration [2], differentiating ECRAM from other non-volatile memory (PCM, RRAM) as synaptic element [3], [4]. However, as previous studies are limited to long time scales, speed and transient effects must be quantified and understood in these devices, and could yet limit their ultimate application. To this end, for the first time, we show that ECRAM can be programmed with a 100 ns pulse and we present subsequent time-resolved data capturing compounded dynamic behavior including transients at scales relevant for device operation.

We measure transients on the order of microseconds and milliseconds in large devices which we attribute to a combination of physical phenomena including RC delay and Li-ion diffusion. We show how transient behavior changes with scaling, and can be further improved through material and device engineering.

2. Experimental

WO_3/LiPON based 3-terminal ECRAM devices were fabricated on patterned and insulated Si substrates using RF sputtering. Figure 1 depicts the device structure. Channel length ranged from 100 μm down to 10 μm.

The device behavior with repeated 1s vertical write pulses is shown in Figure 2. 50+ discrete stable states with long retention times are observed in a resistance range suitable for array implementation. A constant-voltage write scheme results in update asymmetry [5], due to open circuit potential (OCP) build up, as seen in Fig. 2(a). This can be mitigated by material choice or by a constant-current write scheme, as seen in Fig. 2(b). State retention, shown in Fig. 2(c), is maximized by floating the gate after writing.

A test circuit for transient analysis, which allows ns speed current pulses and simultaneous reading of ECRAM gate voltage (V_g) as well as drain current (I_d), is shown in Figure 3. In a functional array, the 2-driver-FET per cell design achieves local gate current pulse control for potentiation and depression, while the gate is left floating during read to maximize retention [6] and the OCP is tracked by the read-out pFET.

Figure 4 shows the transient data following a 100 ns write pulse on a device with 100 μm channel length. Model fits are overlaid in Fig. 4(b-c). We observe non-volatile conductance change on longer timescale preceded by μs and ms transients in read current and slightly slower voltage dynamics at the gate. The latter transient is indicative of the electric field that drives the ions into the channel, and the former will monitor the injected update current from the gate as it reaches the drain in addition to the intended source-drain sense current from changing channel conductance following ion insertion. The charge injected from the gate during programming (1mA) can be more than five orders of magnitude larger than the steady-state (long-term) source-drain read signal (<1 nA @ 100 mV). Therefore, accelerating dissipation of this charge is critical for the detection speed of the updated synaptic weight.

The distributed equivalent circuit device model is shown in Fig. 5(a). The device is sectioned equally along the channel length where the sub-cell behavior for each section was validated by 2-terminal electrochemical impedance spectroscopy (EIS) measurements, Fig. 5(b).

Two effective time constants derived from the model are shown in the Table in Figure 6. τ_{vert}, is the dielectric relaxation time of the electrolyte (modeled by the upper RC) which determines the transfer rate of charge from the surface to the interface capacitor, and is about 35 μs for our samples. τ_{horiz}, accounted for by our distributed model, captures the transit of charge from the center of the channel to the edge drain contact. While τ_{vert} depends only on the electrolyte transport properties, τ_{horiz} is influenced by channel length (L_{ch}) and conductance (G_{ch}).

For device operation, the critical time delay to read an undisturbed state after device update is defined as t_{read}. This requires steady state source-drain sense current to be larger than the transient current through the electrolyte, and therefore depends on the amplitude of the sense current, as well as the slowest of τ_{vert} and τ_{horiz}. For the tested devices, t_{read} is >96 ms and is dominated by τ_{horiz} but can be reduced by increasing the channel conductance (reducing

length) or reducing total injected charge (which scales with channel thickness and length), eventually leading to τ_{vert} as the limiting factor. The effect of channel length is depicted in Fig. 6(b).

The effects of device scaling and material parameters on the time constants are summarized in Figure 7 relative to the measured devices. With scaling to a sub-micron channel dimensions, we anticipate 10,000x reduction in τ_{horiz} and further material and conductivity improvements should enable sub-µs τ_{vert}.

3. Conclusions

High-speed device operation in the ns range is required for successful neuromorphic computing architectures. We demonstrate cell circuitry which allows fast current pulses in a synaptic ECRAM array while also providing state retention. Long-term potentiation and transient measurements under record 100 ns pulses are shown for the first time. For large devices, transient time constants in µs and ms are measured. Using an equivalent circuit model, we demonstrate how the time constants relate to device geometry and material properties, and identify key variables for speed improvement by minimization of such transients.

References

[1] E. J. Fuller, et al. *Adv. Mater.*, vol. 29, (2017) p. 1604310.

[2] T. Gokmen and Y. Vlasov, *Front. Neurosci.*, vol. 10, (2016) 1–13.

[3] G. W. Burr, et al., *IEEE J. Emerg. Sel. Top. Circuits Syst.*, vol. 6, (2016) pp. 146–162.

[4] P. Yao, et al. *Nat. Commun.*, vol. 8, (2017) p. 15199.

[5] S. T. Keene, et al. " *J. Phys. D. Appl. Phys.*, vol. 51, (2018) p. 224002.

[6] S. Kim, et al., *IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS)*, (2017)

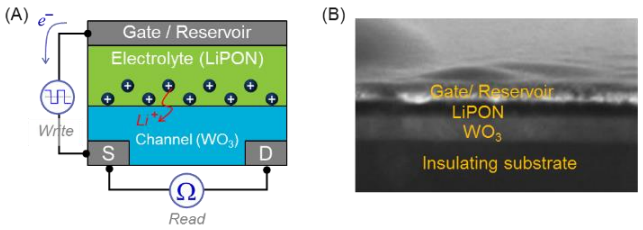

Fig. 1: 3-terminal ECRAM synaptic cell with 10 to 100µm channel length. Programming is achieved by controlling ion motion using the gate terminal. Li-ion intercalation in the channel induces discrete resistance level change sensed by the bottom source-drain contacts. (B) SEM image of channel

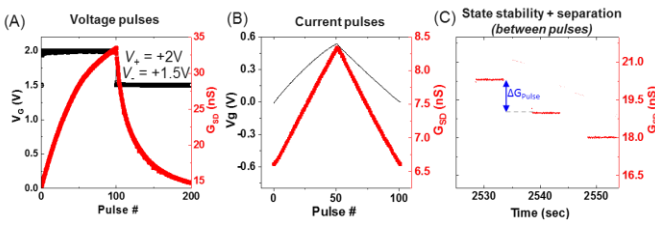

Fig. 2: Device conductance traces with repeated 1sec write. 100mV drain bias was used for reading (A) Voltage-pulse and (B) current pulse operation of the synaptic cell gate. Variation of open circuit potential (OCP) yields asymmetry with V-control and is mitigated with current control. (C) 100+ discrete stable levels are programmed and sensed with the gate floating.

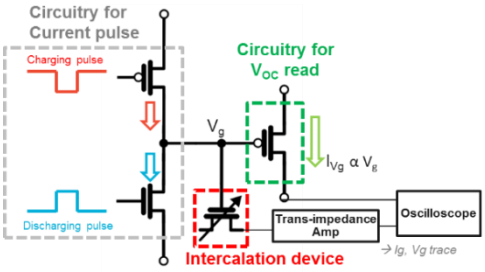

Fig. 3: Test setup for fast current pulse write and fast read. pFET and nFET held at constant Vd are used for charge and discharge respectively and are operated in saturation to supply constant current regardless of Vg. (V_{OC} denotes OCP)

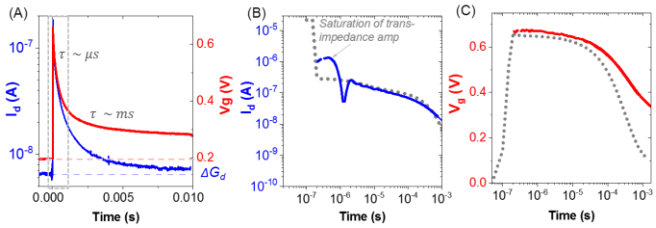

Fig. 4: ECRAM transient response from a single 100 ns write pulse (100 pC equivalent) on a 100 µm channel device. (A) Drain current (I_d) and gate voltage (V_g), showing conductance change. (B) Drain current (log-log) (C) Gate voltage (semi-log). Model overlaid with gray dot.

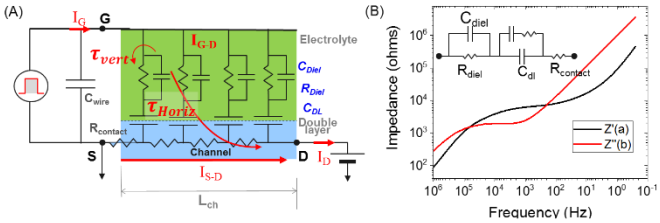

Fig. 5: (A) Model of intercalation ECRAM device showing representation of electrolyte, double layer and channel by resistor-capacitor networks. Ions are transported through the resistive electrolyte and stored in the double layer capacitor where they modulate the conductivity of the highly resistive channel. (B) Real impedance (black) and imaginary impedance (red) from EIS measurement for measuring device parameters with no channel and basic circuit used for EIS fitting.

(A)

Time constant	Scaling relation	Model fit	*Scaled device**
τ_{vert}	$\propto \rho_{diel}\varepsilon_{diel}$	35 µs	3.5 µs
τ_{horiz}	$\propto \frac{\rho_{ch}\varepsilon_{diel}L_{ch}^2}{h_{diel}}$	350 µs	~3.5 ns
t_{read}	$V_dG_d > \frac{\Delta G_d}{\mu_{eff}}\frac{1}{\tau}\exp(-t/\tau)$	~96 ms	~3.5 µs

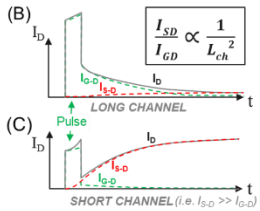

Fig. 6: (A) Table of time constant scaling and model projection. Scaled device based on 100 nm channel and 10x decrease in ρ_{diel}. (B) Illustration of drain current transient for long gate length and (C) short gate length devices showing how the ratio of channel to gate components increases inversely with the gate length (as L_G^{-2})

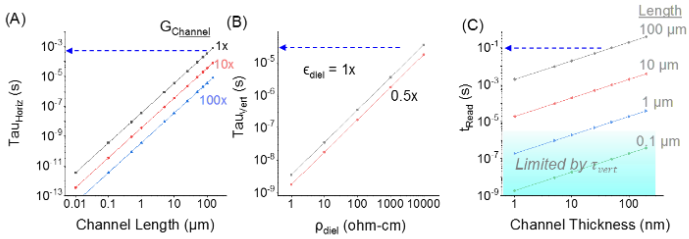

Fig. 7: Scaling limits for time constants in ECRAM relative to measured device (blue arrow). (A) Channel length vs τ_{horiz}, (B) electrolyte ionic conductivity vs τ_{vert}, (C) channel thickness vs t_{read}