ECRAM as Scalable Synaptic Cell for High-Speed, Low-Power Neuromorphic Computing

Jianshi Tang*, Douglas Bishop, Seyoung Kim, Matt Copel, Tayfun Gokmen, Teodor Todorov, SangHoon Shin, Ko-Tao Lee, Paul Solomon, Kevin Chan, Wilfried Haensch, John Rozen
IBM Thomas J. Watson Research Center, Yorktown Heights, New York, USA, *Email: jtang@us.ibm.com

***Abstract*—**We demonstrate a nonvolatile Electro-Chemical Random-Access Memory (ECRAM) based on lithium (Li) ion intercalation in tungsten oxide (WO_3) for high-speed, low-power neuromorphic computing. Symmetric and linear update on the channel conductance is achieved using gate current pulses, where up to 1000 discrete states with large dynamic range and good retention are demonstrated. MNIST simulation based on the experimental data shows an accuracy of 96%. For the first time, high-speed programming with pulse width down to 5 ns and device operation at scales down to 300×300 nm^2 are shown, confirming the technological relevance of ECRAM for neuromorphic array implementation. It is also verified that the conductance change scales linearly with pulse width, amplitude and charge, projecting an ultralow switching energy ~1 fJ for 100×100 nm^2 devices.

I. Introduction

The success of deep learning is related to the availability of large data sets and the use of GPUs to implement the training with the back-propagation algorithm. While different solutions are pursued to increase computing efficiency, the von Neumann bottleneck may eventually prevent further progress. Neuromorphic computing has emerged as a new computing paradigm to enable massively parallel analog computing for deep learning. For example, a new architecture of resistive processing unit (RPU) could provide 30,000× acceleration compared to state-of-the-art CPU/GPU in training deep neural networks [1]. Experimentally, various nonvolatile memories (NVMs), such as resistive random-access memory (ReRAM) and phase-change memory (PCM), have been evaluated as synaptic elements to build prototype neural networks [2-3]. While such NVMs have recently shown encouraging results for inference, their success in training neural network is hampered by their non-ideal switching characteristics, such as asymmetric weight update, stochasticity, and limited endurance.

To circumvent those intrinsic flaws, nonvolatile electrochemical switches have been proposed as artificial synapses for neuromorphic computing [4-5]. As a trade-off for cell complexity using three-terminal device, their read and write operations are decoupled, allowing for better endurance and low-energy switching while maintaining nonvolatility. More importantly, the electrochemically driven intercalation or redox reaction can be precisely and reversibly controlled by the amount of charge through the gate, so they can provide symmetric switching with plentiful discrete states and reduced stochasticity. Indeed, symmetric second-scale switching has been shown on millimeter-size redox transistors with $Li_{1-x}CoO_2$ channel [4]. Meanwhile, organic electrochemical transistors have shown millisecond switching and low switching energy of 10 pJ [5]. However, a clear path to nanosecond switching and device scaling in solid-state electrochemical transistors with sufficient dynamic range remained to be demonstrated to make them technologically relevant. In this work, a nonvolatile ECRAM with up to 1000 discrete conductance levels and large dynamic range is fabricated. Sub-micron devices and high-speed write with pulse down to 5 ns are demonstrated for the first time. An ultralow switching energy ~1 fJ is projected for scaled devices. ECRAM emerges as a promising candidate for high-speed, low-power neuromorphic computing.

II. Switching Characteristics of ECRAM

Fig. 1 illustrates the device structure of ECRAM, where Li ions are electrochemically driven by the gate to (de)intercalate into WO_3 to change its conductance for synaptic weight update. Lithium phosphorous oxynitride (LiPON) is used as a solid-state electrolyte. The amount of Li ions intercalated in WO_3 is precisely controlled by the gate current and this process is reversible, enabling symmetric update. In operation, series of positive (negative) current pulses are fed into the gate for potentiation (depression). As shown in **Fig. 2**, a typical ECRAM is sequentially programmed with 50 up then 50 down pulses (amplitude $I_G = \pm 100$ pA and width $t_w = 5$ s), featuring good symmetry and a large conductance dynamic range ~ 40. The zoom-in view shows discrete conductance states and good retention during read ($t_r = 5$ s) when the gate is floating. It should be noted that the conductance G and its change per pulse ΔG can be tuned by pulse width/amplitude (see **Figs. 10-12** later), device geometry, and material engineering. Such a wide tunability provides advantages over filamentary-type NVMs and previously reported electrochemical switches.

Cycling within a smaller dynamic range achieves a better linearity and symmetry in switching (**Fig. 3**). Here on average 55 up/down pulses ($I_G = \pm 100$ pA and $t_w = 1$ s) are used to cycle G between 1 nS and 3 nS, showing good reproducibility. In **Fig. 4a**, the switching nonlinearity analysis yields nonlinearity of $\nu = 0.347$ and 0.268 for potentiation and depression, respectively, representing near-ideal symmetry and linearity compared to literature (see **Table 1**) [6-7]. The cycle-to-cycle variations from **Fig. 3** (assuming device-to-device variation of $\sigma = 30\%$) is also used in neural network training simulations to yield a more practical classification accuracy. **Fig. 4b** shows the extracted ΔG^+ and ΔG^- per up/down pulse as a function of G over 100 cycles. The asymmetry $AS = |\Delta G^+/\Delta G^-|$ is also plotted

in the inset, varying between 0.6 and 1.6. Building on Ref. [1], a modified weight model that captures ΔG dependency on G along with noise mitigation yields a better trade-off between asymmetry requirement, noise, and number of states. As shown in **Fig. 4c**, the simulated classification accuracy based on our device data (black line) for MNIST dataset is about 96%, close to the ideal numerical floating-point accuracy at 98% (black circle). Increasing the number of states to $N = 110$ (blue line) or reducing variations to $\sigma = 0\%$ (green line) does not affect the result, but the accuracy starts to degrade when N is reduced to 32 (red line). This implies that the accuracy is predominantly determined by the switching asymmetry, which needs to be reduced by about 2× to match the floating-point value as shown in **Fig. 5**. Such reduction in asymmetry is possible by shrinking the dynamic range of G, which can be done at constant N by reducing ΔG with smaller pulses; however, this may increase the noise in the vector-matrix multiplications. Here we use Gaussian noise of $\sigma = 0.06$ as baseline noise [1], and **Fig. 6** shows that MNIST simulations can tolerate up to 7× more noise if appropriate noise management is performed [8]. It should be noted that having access to 1000 levels or more in ECRAM is projected to be significant in larger neural networks for complex datasets beyond MNIST benchmarking.

III. High-Speed Pulse Measurements

To demonstrate high-speed programming of ECRAM, we implement a circuit in which the drain terminals of discrete PFET and NFET are connected at the gate of ECRAM as shown in **Fig. 7.** Here the PFET (NFET) serves as a current source to supply fast current pulses to potentiate (depress) ECRAM by turning on the FETs exclusively. This circuit can be used as a unit cell in an ECRAM-based cross-point array [7]. Using this setup, we demonstrate the first successful sub-μs programming of electrochemical switches. **Fig. 8** shows reproducible cycling through nonvolatile discrete levels with 1 μs pulses with $I_G = \pm 100$ μA. **Fig. 9** shows reproducible cycling with 5 ns pulses with $I_G = \pm 1$ mA. Here the programmed conductance levels are read 1.5 s (pulse period) after each pulse. Write-induced transients that can affect the update frequency and implications of current *vs* voltage operation are discussed elsewhere [9]; material optimization and improved device design are needed to further accelerate read and update in scaled ECRAM. We note that the switching symmetry is slightly degraded compared to **Fig. 3**, due to the non-ideal characteristics of current source FETs: finite output resistance and non-zero drain leakages.

To further understand the programmability of our ECRAM, we perform systematic cycling tests with various pulse widths (t_w) and amplitudes (I_G). **Fig. 10** displays ΔG as a function of t_w down to 5 ns while keeping the identical $I_G = \pm 1$ mA. In **Fig. 11,** we show ΔG as a function of I_G down to ±20 μA with fixed t_w =100 ns. Both trends clearly reveal a linear scaling relation between ΔG and pulse charge Q ($= I_G \times t_w$), which reaffirms the charge-driven nature of ECRAM programming mechanism as shown in **Fig. 12**. Here the data from **Fig. 3** with $t_w \sim 1$ s is also plotted, showing consistency over a wide range of pulse widths. The inset in **Fig. 12** shows that the linear scaling holds down to 2 pC, corresponding to a low switching energy of ~2 pJ per update (assuming an average gate voltage of 1 V). In addition, our ECRAM has shown excellent endurance and no degradation in symmetry across 1000 levels after being cycled with 10^5 pulses, as shown in **Fig. 13**.

IV. Scaling and Projection

To shed light on device scaling, **Fig. 14** demonstrates the switching on a ECRAM device with 1 μm channel length, where much smaller current pulses of $I_G = \pm 1$ pA with $t_w = 5$ s are used for weight update. The scaled device still shows discrete conductance states with good retention and an even larger dynamic range > 10^3. **Fig. 15** further shows switching on a 300×300 nm^2 ECRAM with 100 discrete conductance levels. This demonstrates, for the first time, multi-level operation of ECRAM at scales relevant for large-array implementation. **Fig. 16** shows that the average ΔG scales roughly linearly with the normalized charge by area Q/A. Note that the shortest device channel length on this graph is 500 nm; the smaller devices exhibit an offset ΔG scaling due to geometry effects and distinct patterning. The switching energy E, normalized by ΔG, is plotted as a function of the device size A in **Fig. 17**, which also exhibits linear scaling. It is then extrapolated to yield $E/\Delta G \sim 10^{-4}$ J/S for an ultra-scaled device of 100×100 nm^2 (assuming no geometry effects). Using this number, we can estimate the required current pulse amplitude I_G for a given pulse width t_w to achieve the target ΔG, as shown in **Fig. 18**. For example, given $t_w = 10$ ns, $I_G = 10$ μA is needed to yield ΔG = 1 nS, or just 100 nA for ΔG = 0.01 nS, which corresponds to an ultralow switching energy of about 1 fJ, matching the ultimate energy efficiency of the human brain (~1−10 fJ per synaptic event).

V. Conclusion

In conclusion, we have fabricated a nonvolatile WO_3-based ECRAM that relies on electrochemically driven Li-ion intercalation for neuromorphic computing. Compared to conventional NVMs, ECRAM has shown many unique merits in switching, including superior symmetry and linearity, discrete conductance states with less stochasticity, large dynamic range, and excellent endurance. It is verified that the weight update scales with pulse charge and device size, projecting an ultralow switching energy down to 1 fJ. For the first time, programming with sub-μs pulses and sub-micron devices have both been demonstrated. As summarized in **Table 1**, our work, representing near-ideal symmetry and linearity compared to the literature surveyed in Refs. [6-7], paves the path for using ECRAM in future neuromorphic computing.

Acknowledgment

The authors gratefully acknowledge T.-C. Chen, and Z. Lemnios for executive support; W. Green, V. Narayanan, G. Burr, T. Ando, J. Hannon and J. Tersoff for helpful discussions; and J. Bucchignano and J. Yurkas for technical assistance.

References

[1] Gokmen & Vlasov, *Front. Neurosci.*, **10**, 333 (2016). [2] Burr *et al.*, *IEEE Trans. Electron Devices*, **62**, 3498, (2015). [3] P. Yao *et al.*, *Nat. Commun.*, **8**, 15199, (2017). [4] Fuller *et al.*, *Adv. Mater.*, **29**, 1604310, (2017). [5] van de Burgt *et al.*, *Nat. Mater.*, **16**, 414, (2017). [6] P. Y. Chen *et al.*, *ICCAD,* 194 (2016). [7] Li *et al.*, *VLSI Tech. Dig.*, 25 (2018). [8] Gokmen *et al.*, *Front. Neurosci.* **11**, 538, (2017). [9] Bishop *et al.*, *SSDM*, accepted (2018).

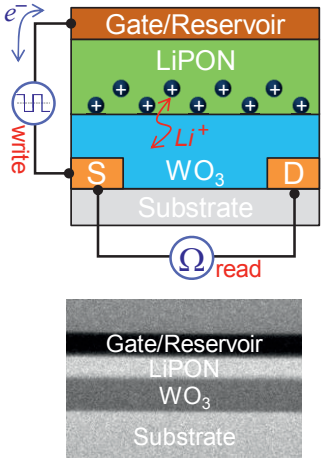

Fig. 1. ECRAM device schematic (top) and cross-sectional TEM image (bottom)

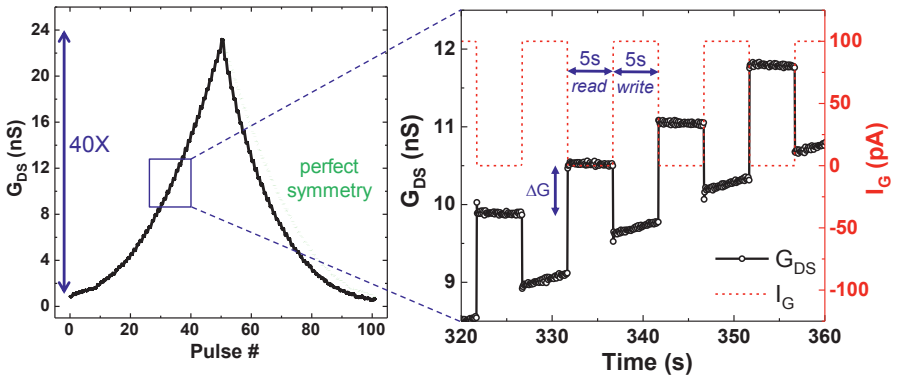

Fig. 2. Plot of the source-drain conductance G_{DS} during gate current pulses (50 up then 50 down pulses with amplitudes $I_G = \pm 100$ pA and width $t_w = 5$ s), showing good symmetry and a large on/off ratio ~ 40. This device has a channel size of $L \times W = 10 \times 60$ μm^2. The green dotted line illustrates the reflection of up trace as a guide to the eye. The zoom-in view (right) shows discrete conductance states with good retention during read. A constant source-drain bias of $V_{DS} = 0.1$ V is applied to monitor the channel conductance during read and write.

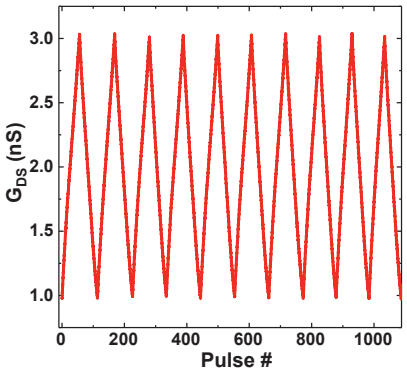

Fig. 3. Reproducible cycling demonstrates both good symmetry and linearity within a conductance range of 1−3 nS. Each cycle has 55 up/down pulses on average with $I_G = \pm 100$ pA and $t_w = 1$ s. This device has a channel size of 80×100 μm^2.

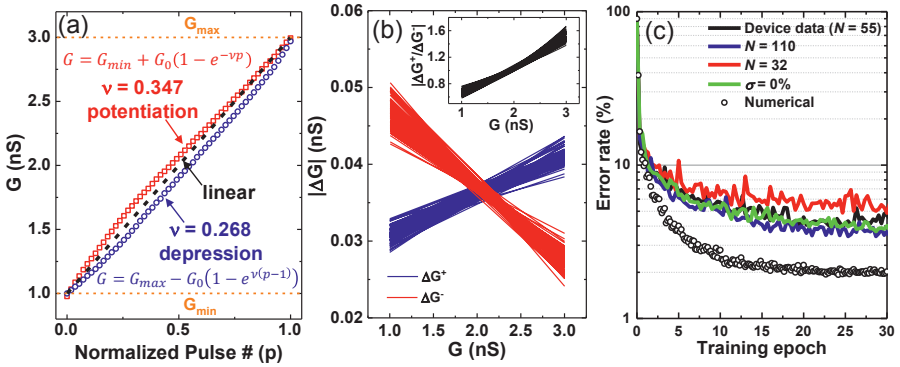

Fig. 4. (a) Nonlinearity analysis on the switching characteristics yields small nonlinearity factors of ν = 0.347 and 0.268 for potentiation and depression, respectively. (b) Plots of ΔG per up/down pulse and asymmetry $|\Delta G^+/\Delta G^-|$ (inset) as a function of G. (c) Taking the cycle-to-cycle variation into consideration, MNIST simulation (assuming another device-to-device variation of $\sigma = 30\%$) shows an accuracy of about 96%, approaching the ideal numerical accuracy. The accuracy is not affected when increasing the number of states ($N = 110$) and reducing variations ($\sigma = 0\%$), but it starts to degrade when N is reduced to 32.

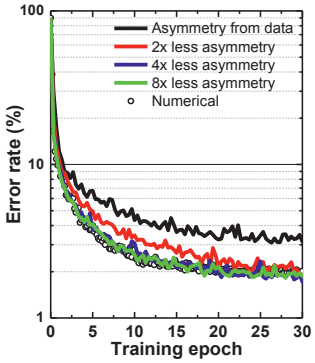

Fig. 5. MNIST simulation accuracy shows clear dependence on the switching asymmetry, which needs to be reduced by just 2× to match the ideal numerical accuracy.

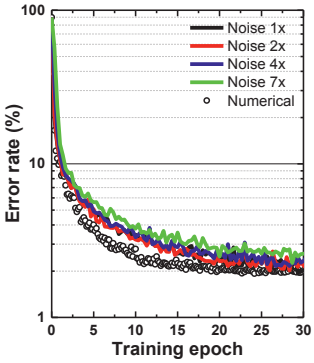

Fig. 6. The effect of noise on the MNIST simulation accuracy, which is shown to be able tolerate up to 7× of the Gaussian noise specified by RPU requirement.

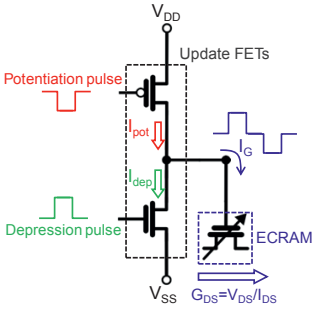

Fig. 7. ECRAM unit cell design for high-speed programming, in which discrete PFET and NFET serve as current source to ECRAM for positive (potentiation) and negative (depression) weight update, respectively.

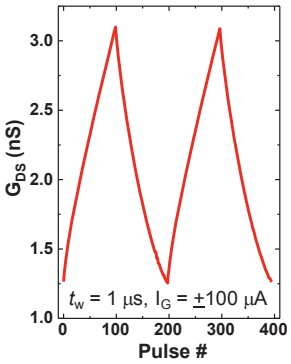

Fig. 8. Reproducible cycling through nonvolatile discrete levels with 1 μs pulses with amplitude of ±100 μA. The pulse period is 1.5 s.

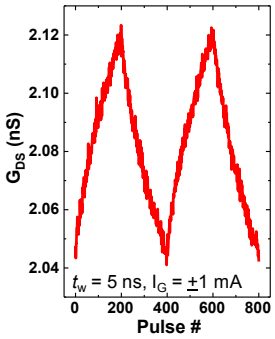

Fig. 9. Reproducible cycling with 5 ns pulses with amplitude of ±1 mA. The pulse period is 1.5 s. The slightly degraded switching symmetry is due to the non-ideal characteristics of current source FETs.

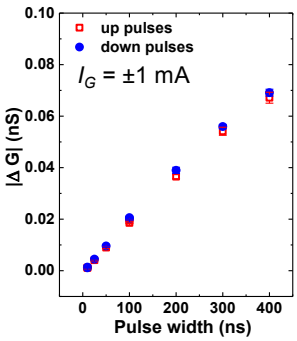

Fig. 10. The average change in conductance ΔG is shown to scale linearly with pulse width from 400 ns down to 5 ns while keeping the same I_G = ±1 mA.

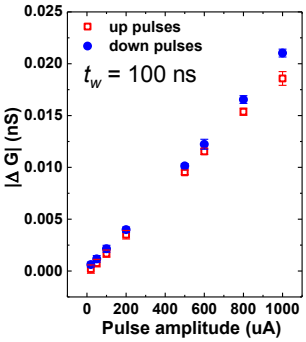

Fig. 11. The average change in conductance ΔG also scales linearly with pulse amplitude from 1 mA down to 20 μA while keeping the identical t_w = 100 ns.

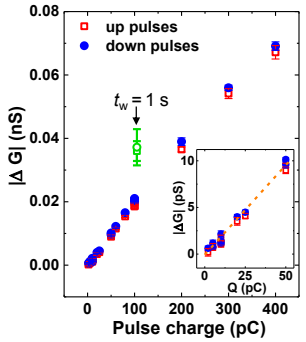

Fig. 12. ΔG scales with pulse charge Q (= $I_G \times t_w$). The two green data points (I_G = ±100 pA, t_w = 1 s) are from Fig. 3. The "zoom-in" inset shows a linear scaling down to 2 pC per pulse.

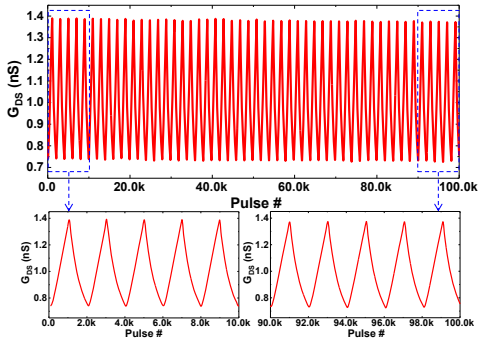

Fig. 13. Endurance test on an ECRAM with 10^5 pulses (I_G = ±100 μA, t_w = 100 ns), showing no degradation in symmetry. There are 10^3 up/down pulses in each cycle while reading the device conductance after every 10^2 pulses.

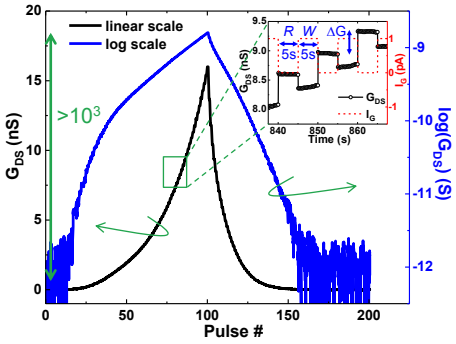

Fig. 14. Demonstration of switching on a scaled ECRAM (1×10 μm²) using ±1 pA current pulses (pulse width t_w = 5 s) for positive/negative weight update. The scaled device still shows good retention and an even larger on/off ratio > 10^3.

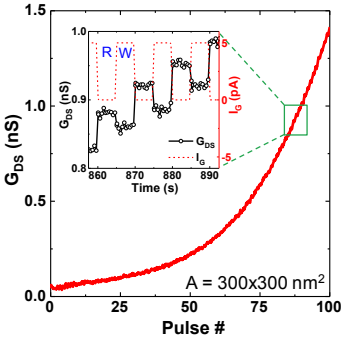

Fig. 15. Programming of a 300×300 nm² ECRAM with 100 up pulses (I_G = 5 pA, t_w = 5 s), showing multi-level operations at scale. Here the ΔG *vs* A scales differently due to geometry effects.

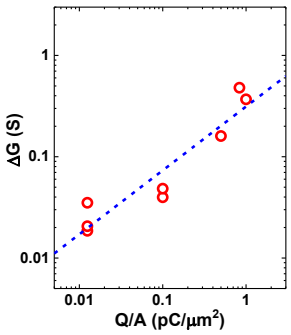

Fig. 16. Plot of the change in conductance ΔG versus pulse charge per device area Q/A, showing a roughly linear scaling.

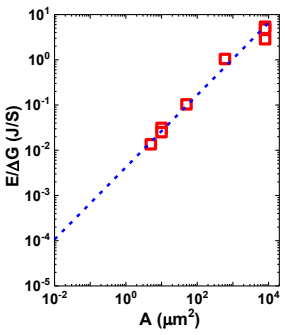

Fig. 17. The switching energy per conductance change $E/\Delta G$ scales linearly with device area A (assuming an average gate voltage ~ 1V).

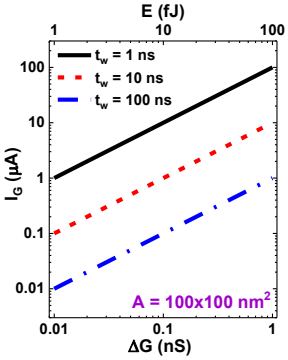

Fig. 18. Projection of the required current pulse amplitudes I_G for given pulse widths t_w in a 100×100 nm² device, where the switching energy can be as low as 1 fJ to make a weight update of ΔG = 0.01 nS.

on/off ratio	40–10^3
# of states	1000 (tunable)
G range	0–24 nS (tunable)
$\lvert\Delta G^+/\Delta G^-\rvert$	0.6–1.6
Nonlinearity ν	0.347 / 0.268
Write pulse width	down to 5 ns
Smallest channel size	0.09 μm²

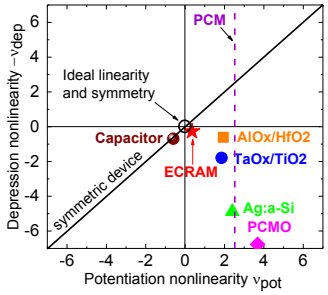

Table 1. Summary of ECRAM key metrics for neuromorphic computing and comparison with other technologies surveyed in Ref. [6-7]. The sign of ν_{dep} is corrected for consistency with literature.