Metal-oxide based, CMOS-compatible ECRAM for Deep Learning Accelerator

Seyoung Kim*, Teodor Todorov, Murat Onen, Tayfun Gokmen, Douglas Bishop, Paul Solomon, Ko-Tao Lee, Matt Copel, Damon B. Farmer, John A. Ott, Takashi Ando, Hiroyuki Miyazoe, Vijay Narayanan and John Rozen
IBM Thomas J. Watson Research Center, Yorktown Heights, New York, USA, *Email: sykim@us.ibm.com

***Abstract*—** We demonstrate a CMOS-compatible, metal-oxide based Electro-Chemical Random-Access Memory (MO-ECRAM) for high-speed, low-power neuromorphic computing. The device demonstrates symmetric and linear conductance update, large on/off ratio and good retention while also withstanding high temperature treatments necessary for BEOL compatibility. Resistive switching in MO-ECRAM is observed with voltage pulses down to 10 ns and scales exponentially with voltage pulse amplitude, enabling parallel array operations without any selector/access devices. For the first time, we experimentally demonstrate fundamental techniques for fully-parallel array operations, stochastic update scheme and zero-shifting technique, and show a successful stochastic gradient descent algorithm demonstration in hardware using a MO-ECRAM array.

I. Introduction

Cross-point arrays of novel resistive memories have been proposed as an alternative computing paradigm to accelerate matrix operations for neural networks (NN) in deep learning applications. Studies have shown that significant acceleration is achievable by utilizing massive parallelism in such analog accelerators [1]. To date, a variety of nonvolatile memory devices (NVMs), including resistive random-access memory (RRAM) and phase-change memory (PCM) [2], have been evaluated as synaptic elements and used to build neural network prototypes. While rapid progress has been shown recently, non-ideal switching characteristics of these devices including asymmetric weight update, cycle-to-cycle and device-to-device variations, and stochasticity are yet to be solved in order to achieve the promised acceleration for neural network training applications.

Recently three-terminal, nonvolatile electrochemical switches, i.e. Li-based ECRAM [3-5], have been proposed as synaptic devices for neuromorphic computing, featuring superior switching properties: symmetric switching response, fast switching speed, low programming energy, and a large dynamic range. Unlike their two-terminal counterparts, the three-terminal devices allow separation of read and write operations which leads to better control over the channel modulation. However, typical materials to build ECRAM devices are not fab-friendly, which creates significant obstacles to process integration. Also, due to the barrier-less gate discharge characteristics, an additional access device or selector transistors are required to isolate the device terminals and prevent leakage current [3, 4]. In this work, a metal-oxide based ECRAM is fabricated with all CMOS-compatible materials. Resistive switching is preserved after annealing at 400˚C, and high-speed write with pulses as short as 10 ns are demonstrated. Furthermore, the non-linear gate characteristics precludes the need for selectors or access devices in an array. Training demonstrations on these MO-ECRAM arrays illustrate their programmability, retention and endurance along with the novel techniques for fully-parallel array operations. Our results show that MO-ECRAM is a technologically relevant candidate for high-speed, low-power deep learning accelerators.

II. Switching Characteristics of MO-ECRAM

Fig. 1 illustrates the device structure of MO-ECRAM, where three different metal oxide layers are used to form a switching channel and gate dielectric of a 3-terminal synaptic memory device. The gate terminal is used to apply an electric field to the channel, induce electrochemical changes in the system, and consequently modulate the conductance of the WO_3 channel. The conductance of this channel can be read by applying a voltage between source and drain (V_{ds}) and measuring the channel current. Channel dimensions investigated in this work range from W/L = 100/100 um down to 10/4 um while sub-um scaling has been achieved using standard patterning methods such as reactive-ion etching and lithography thanks to the CMOS-compatibility.

Programmability of MO-ECRAM is tested by providing a series of current pulses to the gate electrode for positive (up) and negative (down) conductance update, while the channel conductance G is measured at V_{ds} = 10 mV. Near-ideal symmetry and linear conductance modulation with a dynamic range of ~8× is demonstrated in **Fig. 2a, c** with 20 up and 20 down pulses of I_G = ±1 nA and a programming pulse width of t_{prog} = 0.5 s. The conductance evolution as a function of time in **Fig. 2b** reveals switching between multiple discrete conductance steps with a uniform ΔG and good retention during the read (t_{read} = 1 s) when the gate is floating (I_G = 0 A). We note that the observed dynamic range is subject to the pulsing conditions, and the maximum conductance contrast can be over ~10^4 in typical MO-ECRAM devices. BEOL thermal budget compatibility is verified by switching tests after high temperature annealing. **Fig. 3** shows that the device maintains reproducible switching over 20 cycles with excellent symmetry using unconditioned, open-loop pulse measurement after annealing 400˚C for 1 min in N_2 environment. Further annealing studies confirmed resistive switching in MO-ECRAM after 1 hour annealing at 400˚C.

High-speed switching of synaptic memory devices is a key requirement to implementing a fast neuromorphic computation system. We demonstrate conductance modulation with voltage pulses widths from 10 µs to 100 ns (**Fig.4a**) at V_{pulse} = ±4 V. At 10 ns pulse width, resistive switching with a small on/off ratio is observed in **Fig. 4b**, confirming that conductance modulation is possible at this timescale. We note that the actual pulse width applied to the device by the source transistors [3] can be longer due to the circuit parasitics. Follow-up study is required to further investigate the ultimate speed of MO-ECRAM in view of the µs-scale read transients observed in Li-ECRAM [5].

Further analysis reveals that in MO-ECRAM, ΔG is logarithmically dependent on pulse width (**Fig. 5a**), exponentially dependent on voltage pulse amplitude (**Fig. 5b**), and linearly related to current pulse amplitude (**Fig. 5c**). Unlike Li-ECRAM where ΔG is proportional to the amount of charge injected [3], the dependence of ΔG on pulse width in MO-ECRAM is found to be non-linear. The cause of this observation requires further studies. The current and voltage dependences of ΔG is closely related to the exponential I-V characteristics of metal oxide stack and supports that both current- and voltage-based programming are possible in this device. Since voltage-based programming allows array hardware without access/selector devices, we choose to use voltage pulses in the array demonstration.

MNIST simulations are performed by extracting the relevant device parameters from the MO-ECRAM to evaluate neural network training performance. Results shown in **Fig. 6** confirm that the achieved symmetry is very close to the ideal level. Therefore, as the number of states is increased to the required level [1], MO-ECRAM can match the training performance of the ideal cross-point element

Retention characteristics of a MO-ECRAM after resistive switching is measured in **Fig. 7**. The conductance value stayed within ~±5% for >14 hours. We note that no change in retention behavior was observed when the gate terminal is either grounded or floating. Strong endurance characteristics is demonstrated in **Fig. 8** by repeatedly switching the device with over 20 million pulses. To understand the energy consumption, we plot switching energy normalized by unit conductance change (E/ΔG) as a function of device area in **Fig. 9**. E/ΔG decreases linearly as a function of area. While follow-up studies are required to confirm this trend at scaled devices, this result indicates a promising trend of switching energy reduction at smaller devices.

III. Array-Level Demonstrations

To verify the interoperability of MO-ECRAM devices in a cross-point array, we connect four discrete devices into a 2×2 array configuration, where drains (sources) of the devices in one row (column) are connected to each other, respectively, as shown in **Fig. 10a**. Because of the exponential gate characteristic of the MO-ECRAM, gate terminals in the same row can also be directly connected to each other and used for update operation without cross-talk. By applying voltage pulses with opposite polarities at matching gate lines and columns using the half-voltage selection scheme, uniform and reproducible programming is observed at both parallel and sequential programming tests, with minimal disturbance at unselected devices (**Fig. 10c**). Since no FEOL elements or additional selectors are required, MO-ECRAM arrays can be integrated at BEOL, on top of other FEOL circuit components such as integrators and analog-to-digital converters. This is a major advantage in implementing area-efficient neuromorphic processors.

With a MO-ECRAM cross-point array, we experimentally demonstrate linear regression by implementing a stochastic gradient descent algorithm and using two key techniques required for resistive memory-based NN accelerators: (i) the zero-shifting technique [6] and (ii) the stochastic weight update scheme [1]. Zero-shifting significantly improves NN performance by restoring regularization terms, impaired by a mismatch between the weight device *symmetry point* (G_{sym}, where ΔG for up and down update are equal) and the reference device conductance (G_{ref}). As illustrated in **Fig. 11a**, zero-shifting requires the use of two resistive memory device arrays as a differential pair to represent one weight matrix. We find (*symmetry point measurement*, see **Fig. 11b**) and copy the symmetry point conductance values from the weight array into the reference array to complete zero-shifting.

Once the reference array is programmed, the weight array is trained to perform linear regression. First, we generate a dataset (x_i, y_i) with intentional Gaussian noise for training (**Fig. 12a**). Then, for each input $\boldsymbol{x_i} = [x_i, 1]$, the multiplication between the input vector and weight matrix is compared with the target value y to generate the error vector $\boldsymbol{\delta_i} = y_i - ([x_i, 1] * [w_{11}, w_{12}]^T)$. Finally, via the stochastic update scheme, vectors $\boldsymbol{x_i}$, $\boldsymbol{\delta_i}$ are translated into probabilistically populated pulse streams (**Fig. 10b**) to implicitly compute their outer product and perform weight update correspondingly [7]. By repeating this procedure, we can train the MO-ECRAM array and solve the linear regression problem. **Fig. 12** shows the evolution of weights throughout the training, where the weights are tuned in parallel to their target values by the back-propagation algorithm. Here, we intentionally set a low learning rate to test the robustness of switching. After over 24 hours of training time and 300k update pulses applied to the weight devices, a successful convergence to the target values is observed, evincing the excellent programmability, retention and endurance of MO-ECRAM.

We note that the stochastic update scheme is essential to program a N×N array in parallel using 2N terminals. The key advantages of stochastic update scheme are accuracy in update and flexibility. The stochastic update scheme can generate update errors at each update cycle due to its random nature, however, the accumulated updates will be accurate on average. On the other hand, deterministic (and parallel) update scheme either makes the same update errors due to the limited resolution or requires a longer bit length (i.e. number of pulses generated per vector) to accomplish the long-term accuracy. Other methods which require explicit computation of the outer product and/or serial programming severely reduce the operation speed. These considerations further signify the importance of the first stochastic update scheme-based training demonstration without using any access or compliance circuitry.

IV. Conclusion

In conclusion, we have demonstrated a CMOS-compatible, nonvolatile, metal-oxide based ECRAM and its array operations with many promising metrics (**Table. 1**) for future neuromorphic computing. Follow-up studies on read/update transients and optimization of the channel conductance range to match the requirements for large cross-point arrays will pave the path for enabling analog memory device-based deep learning accelerators.

Acknowledgment

The authors gratefully acknowledge T.-C. Chen and M. Khare for executive support; W. Haensch, M. Rasch and J. Hannon for insightful discussions.

References

[1] Gokmen & Vlasov, *Front. Neurosci.*, **10**, 333 (2016). [2] Burr *et al.*, *IEEE Trans. Electron Devices*, **62**, 3498, (2015). [3] Tang *et al.,* IEDM, (2018). [4] Fuller *et al.,* Science, **364**, 570-574, (2019). [5] Bishop *et al.*, SSDM, pp23-24, (2018). [6] Kim *et al.*, arXiv:1907.10228, (2019). [7] Gokmen *et al.*, *Front. Neurosci.* **11**, 538, (2017).

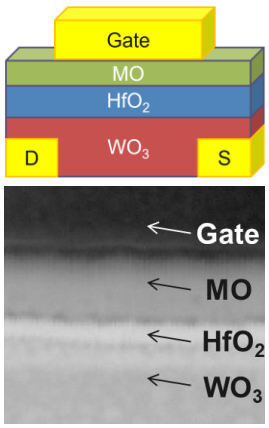

Fig. 1. MO-based ECRAM device schematic (top) and cross-sectional TEM image (bottom)

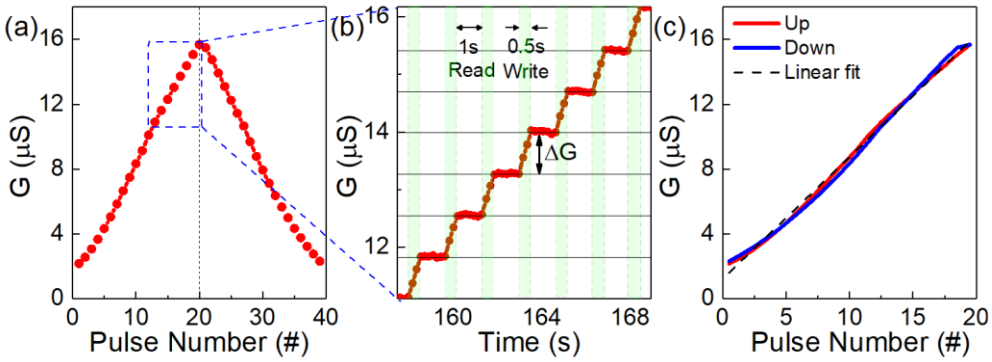

Fig. 2. (a) Conductance (G) modulation by applying current pulses at the gate (20 up then 20 down pulses with $I_G = \pm 1$ nA and pulse width $t_{prog} = 0.5$ s) after 400˚C annealing in N_2 ambient for 1 min. V_{DS} = 10mV is applied to measure the conductance. ~8× of conductance contrast is observed which is dependent on pulse conditions. (b) The zoom-in view reveals the discrete conductance states with a uniform conductance difference. Read time of 1s is used to monitor the conductance between the programming pulses with $I_G = 0$ A, and good retention is observed during the read. (c) The up and down switching characteristics show excellent symmetry and linearity.

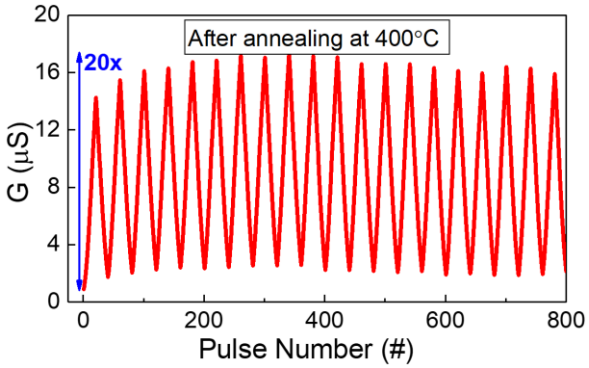

Fig. 3. Reproducible cycling demonstration for 20 cycles with no symmetry degradation and ~20× conductance contrast. Each cycle is 20 up/down pulses with $I_G = \pm 1$ nA and t_{prog} = 0.5 s. The measurement is performed with open-loop without any external control.

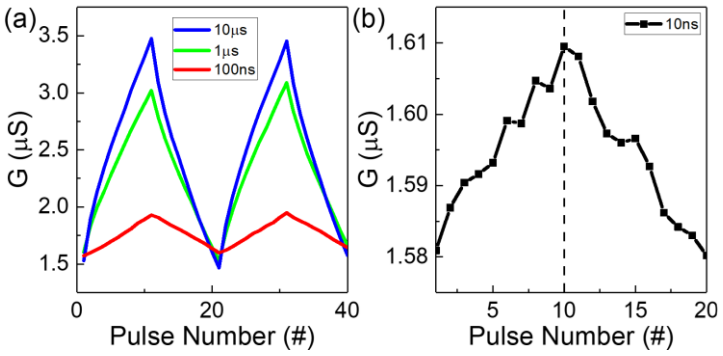

Fig. 4. (a) High-speed switching of MO-ECRAM is verified by applying 10 up/down voltage pulses ($V_{GS} = \pm 4$ V) with 10 μs, 1 μs and 100 ns pulse widths for 2 cycles. More symmetric switching is observed as pulse width decreases. (b) Conductance switching in a MO-ECRAM device (W/L = 20/80 um) is demonstrated with a pulse width of 10 ns.

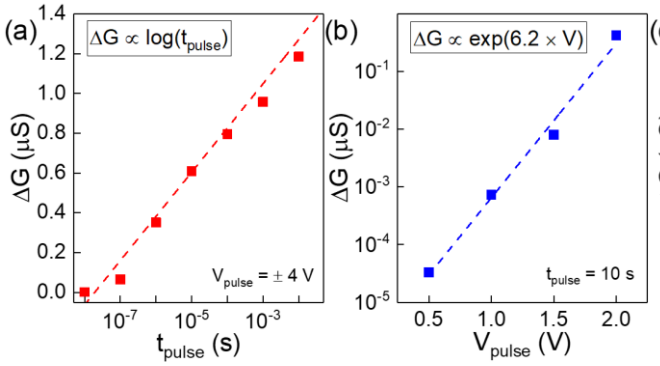

Fig. 5. (a) Change in the channel conductance (ΔG) versus pulse time given at the instrument during voltage pulse programming at $V_{GS} = \pm 4$ V and with a read period of 0.2 s. The conductance modulation shows logarithmic dependence to pulse time. We note that when multiple pulses with a fixed pulse width are supplied with a time interval, ΔG will be added up linearly, enabling cumulative updates. (b) ΔG versus programming pulse amplitude. The programming voltage dependence is exponential which leads to linear current dependence due to exponential I_G-V_{GS} characteristics. (c) ΔG versus average current supplied during voltage pulse programming. ΔG closely follows a linear relation with the current. Both up and down pulse data are included in the analysis for all 3 panels.

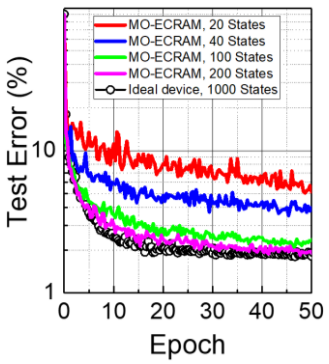

Fig.6. MNIST simulation accuracy based on characteristics shown in Fig. 2. If we increase the number of states from 20 to 40, 100 and 200, MO-ECRAM shows neural network accuracy converging to that of ideal symmetric devices.

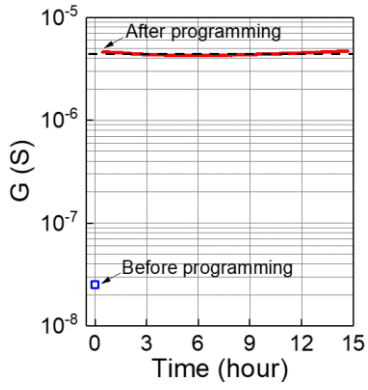

Fig. 7. Retention measurement for >14 hours after programming. Maximum ~5.4 % conductance drop is observed at higher conductance states.

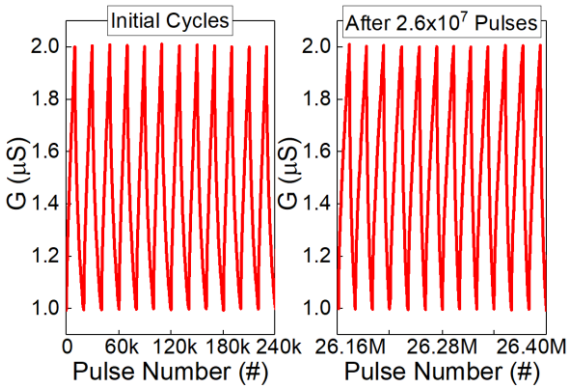

Fig. 8. Demonstration of endurance for 2.6 × 10^7 pulses on a MO-ECRAM using a conditioned cycling measurement with on/off ratio of 2. Resistive switching is observed after ~1300 cycles of 10,000 up/10,000 down pulses on average. The device is still working after the last cycle of this measurement.

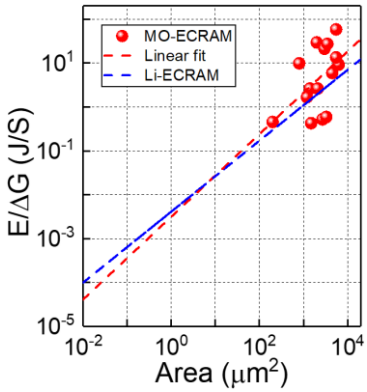

Fig. 9. Switching energy per unit conductance change, E/dG, scales linearly with device area A. MO-ECRAM shows similar scaling trend with that of Li-based ECRAM [3].

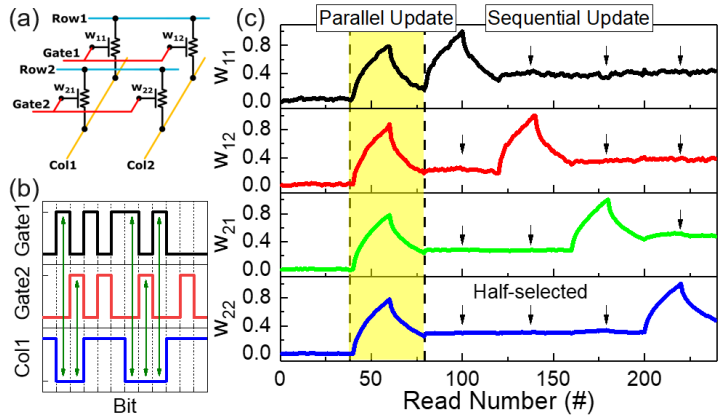

Fig. 10. Interoperability of MO-ECRAMs in a 2×2 cross-point array configuration is verified without any selector/access device. (a) 2x2 array of MO-ECRAM devices. (b) An example set of stochastic pulses for parallel update. w_{11} sees three updates while w_{21} sees two update pulses. (c) All four devices show uniform, reproducible switching when selected in parallel or selected individually. Minimal disturbance at unselected devices is observed with half-voltage selection scheme.

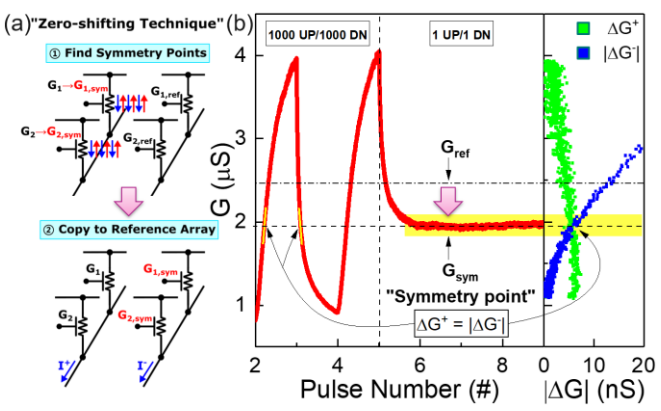

Fig. 11. (a) Illustration of the zero-shifting technique and (b) symmetry point measurement demonstration. Conductance is monitored for multiple cycles of 1000 up/1000 down pulses (left) to obtain the G versus ΔG relation in the conductance range (right). Then, 2000 cycles of 1 up/1 down pulses are applied to converge the conductance to the symmetry point, which matches with the equal |ΔG| points at 1000 up/1000 down measurement.

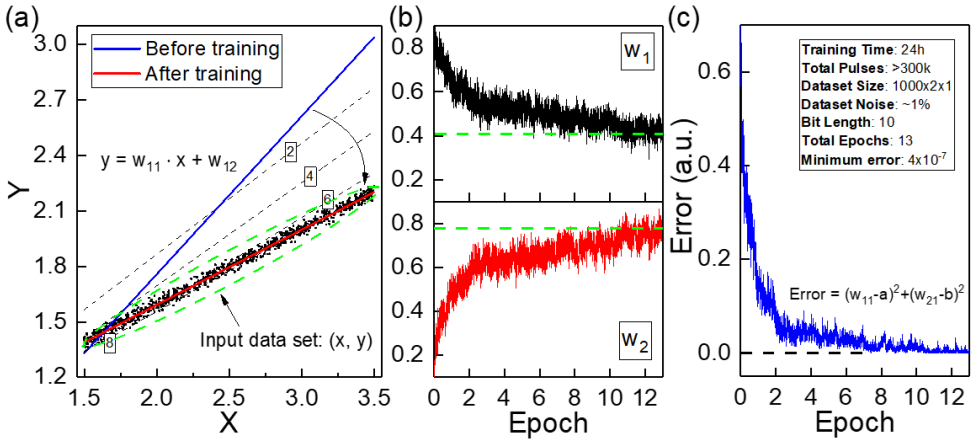

Fig. 12. First experimental demonstration of stochastic gradient descent algorithm using stochastic update scheme in a 2×2 MO-ECRAM array. As training proceeds, (a) the fitting line converges to the input data set, (b) w_{11} and w_{21} are approaching to the target values, and (c) the error converges to zero. Inset of (c) lists parameters and relevant numbers used in this demonstration. 1,000 input data points are trained at each epoch. To incorporate the effect of retention, the learning rate was purposefully set to be low, leading to a training time of 24 hours.

Property\Device	Li-ECRAM	MO-ECRAM
CMOS/BEOL	Incompatible	Compatible
Access Circuitry	Required	Not Required
Asymmetry, $\left\|\frac{\Delta G}{\Delta G^-}\right\|$	0 6/1 6	1 005/0 995
# of states	1000 (tunable)	1000 (tunable)
G range	0 – 24 nS (tunable)	0 – 50 μS (tunable)
Write pulse width (transients to be evaluated [5])	≤ 5 ns	≤ 10 ns
E/ΔG (*100x100nm², projection)	~*100 fJ/nS	~*100 fJ/nS

Table 1. Summary of MO-ECRAM key metrics for neuromorphic computing and comparison with Li-ECRAM technology [3].