Artificial Synapse based on ULTRARAM Memory Device for Neuromorphic Applications

Abhishek Kumar*,[1] Peter D. Hodgson,[2,3] Manus Hayne,[2,3] and Avirup Dasgupta*,[4]
[1]*Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, CA 94720, USA*
[2]*Department of Physics, Lancaster University, Lancaster LA1 4YB, United Kingdom.*
[3]*Quinas Technology Limited, Lancaster LA1 4YB, United Kingdom.*
[4]*Department of Electronics and Communication Engineering, Indian Institute of Technology Roorkee, Roorkee 247667, India.*

(*Electronic mail: abhishekg@berkeley.edu, avirup@ece.iitr.ac.in)

The memory demands of large-scale deep neural networks (DNNs) require synaptic weight values to be stored and updated in off-chip memory like dynamic random-access memory, which reduces energy efficiency and increases training time. Monolithic crossbar or pseudo-crossbar arrays using analog non-volatile memories, which can store and update weights on-chip, present an opportunity to efficiently accelerate DNN training. In this article, we present on-chip training and inference of a neural network using an ULTRA**RAM** memory device-based synaptic array and complementary metal-oxide-semiconductor (CMOS) peripheral circuits. ULTRA**RAM** is a promising emerging memory exhibiting high endurance ($>10^7$ P/E cycles), ultra-high retention (>1000 years), and ultra-low switching energy per unit area. A physics-based compact model of ULTRA**RAM** memory device has been proposed to capture the real-time trapping/de-trapping of charges in the floating gate (FG) and utilized for the synapse simulations. A circuit-level macro-model is employed to evaluate and benchmark the on-chip learning performance in terms of area, latency, energy, and accuracy of an ULTRA**RAM** synaptic core. In comparison to CMOS-based design, it demonstrates an overall improvement in area and energy by 1.8× and 1.52×, respectively, with 91% of training accuracy.

I. INTRODUCTION

Deep neural networks (DNNs) have demonstrated remarkable success across various applications, including image classification, speech recognition, time-series prediction, and spatiotemporal recognition tasks[1,2]. However, DNNs implemented on conventional von Neumann computing architectures suffer from significant energy consumption and high latency[3]. This is due to the memory demands of the large-scale neural networks often surpassing the capacity of on-chip SRAM caches[4]. Additionally, expanding SRAM size is constrained due to the considerable cell area requirement (100-200F^2), making scalability a challenge[5,6]. As a result, high-bandwidth off-chip memory, such as DRAM, is commonly utilized to store network parameters[7]. However, this approach reduces energy efficiency and increases latency compared to on-chip solutions due to the constraints of the von-Neumann bottleneck[8,9]. In a fully connected DNN, training can be significantly accelerated by reducing data movement through on-chip storage and conducting weight updates directly at the same node, with all nodes interconnected within an array.

Monolithic crossbar or pseudo-crossbar arrays using analog non-volatile memories, which can store and update weights on-chip, present an opportunity to accelerate DNN training by reducing data movement[10]. Various emerging non-volatile memory technologies, such as resistive random-access memory (RRAM)[11,12], phase-change memory (PCM)[13], and ferroelectric devices[14,15], are promising candidates due to their compact cell size and capability to store multiple intermediate states. However, PCM experiences a sudden reset transition, whereas oxygen vacancy-based RRAM devices are prone to cycle-to-cycle variability and limited G_{max}/G_{min} ratios, which leads to asymmetric potentiation and depression characteristics[16]. Additionally, the slow write speeds, ranging from microseconds to milliseconds, can significantly prolong training duration, potentially extending to several years[14,17].

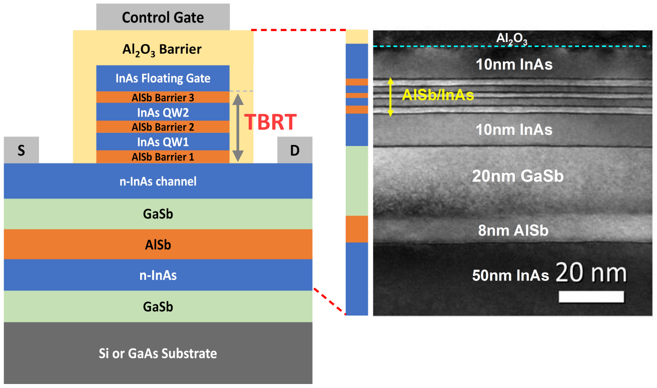


FIG. 1: Schematic of an ULTRA**RAM** memory cell and the corresponding transmission electron microscope image of the device's epilayers.[18] From Lane et al., IEEE Trans. Electron Dev. 68(5), 2271–2274 (2021). Copyright 2021 Author(s), licensed under a Creative Commons Attribution 4.0 (CC BY 4.0) license.

In this paper, we present on-chip training and inference of a neural network using an ULTRA**RAM** memory device-based synaptic array and CMOS peripheral circuits. A physics-based compact model of an ULTRA**RAM** memory device has been used to capture the real-time trapping/de-trapping of charges in the floating gate (FG) and utilized for the synapse[19,20]. A circuit-level macro-model is employed to evaluate and benchmark the on-chip learning performance in terms of area, latency, energy, and accuracy of an UL-

TRA**RAM** synaptic core[21]. In comparison to CMOS-based SRAM design, it demonstrates an overall improvement in area, energy, and latency with 91% training accuracy.

II. MEMORY PROPERTIES AND MODELING

ULTRA**RAM** is a promising emerging memory exhibiting high endurance (>>10^7 P/E cycles[22]), ultra-high retention (>1000 years), and ultra-low switching energy per unit area[18,23]. The state is determined by the presence or absence of electrons in a floating gate (FG). Unlike a single SiO_2 barrier in flash memory, the novelty comes from the InAs/AlSb triple barrier resonant tunneling (TBRT) structure[24], as shown in Fig. 1. TBRT structure provides a high-potential electron barrier with no bias and allows fast resonant tunneling to program/erase pulse (±2.5V) with switching energy per unit area 1000 times lower than NAND flash, and 100 times lower than DRAM[25]. The ULTRA**RAM** cells were simulated using a physics-based compact model that self-consistently links resonant tunneling through the triple-barrier stack with floating-gate charge storage and channel conduction[19,20]. The tunneling current through the TBRT structure is described using an energy-resolved resonant tunneling formulation, where the current density is obtained by integrating the transmission probability over the longitudinal carrier energy distribution as follows

$$J_i = \frac{q_e m^* kT}{2\pi^2 \hbar^3} \int_0^\infty T(E_x,V) ln\left[\frac{1+exp(\frac{E_f-E_x}{kT})}{1+exp(\frac{E_f-E_x-q_eV}{kT})}\right] dE_x, \tag{1}$$

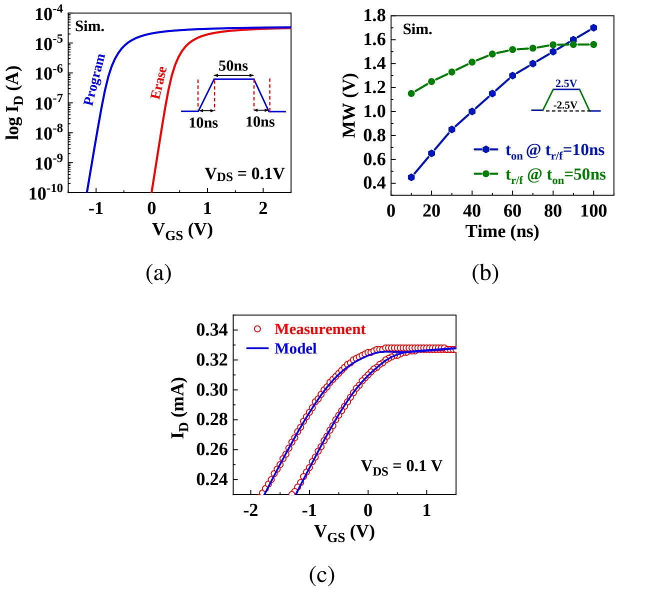
where $T(E_x,V)$ is the voltage-dependent transmission coefficient, q_e is the charge of an electron, m^* is the effective mass of the electron, k is the Boltzmann's constant, T is the absolute temperature, $\hbar$ is the reduced Planck's constant, E_x is the longitudinal energy, V is the potential applied to the structure, and the term with the log function represents the carrier supply function determined by Fermi-Dirac statistics. Each quantum well resonance is modeled using a Lorentzian transmission profile centered at the bias-shifted resonance energy, enabling accurate reproduction of the sharp current peaks during program and erase operations[19]. For the ULTRA**RAM** stack, contributions from multiple resonant levels are summed, $J_{TBRT} = \sum_{i=1}^{n-1} J_i + J_{th}$, where n is the number of barriers, with an optional empirical thermionic term included as $J_{th} = H(exp(q_eV/2kT)-1)$ to account for high-field transport when required.

The tunneling current density through TBRT (J_{TBRT}) is dynamically integrated to obtain the floating-gate charge (Q_{FG}),

$$\frac{dQ_{FG}}{dt} = A\, J_{TBRT}, \tag{2}$$

where A is the effective tunneling area. This time-dependent floating-gate charge shifts the effective gate voltage ($V_{gs,eff}$), and hence the device threshold voltage. The drain current during read operation is calculated using a surface-potential-based channel model[26],

$$I_{ds} = \mu_{eff} C_g \frac{W}{L}\left(V_{gs,eff} - V_{off} - \frac{Q_{FG}}{C_g} - \psi_m\right)\psi_{ds}, \tag{3}$$



(a)

(b)

(c)

FIG. 2: (a) Simulated transfer characteristics of ULTRA**RAM** at W = L = 1 μm. (b) Variations in the memory window (MW) of the ULTRA**RAM** cells with applied input pulse width and rise/fall time. (c) Validation of the model with experimental long-channel (L = 10μm) I-V characteristics[23].

where C_g is the gate capacitance, μ_{eff} is the effective mobility, V_{off} is the cut-off voltage, and ψ_m and ψ_{DS} are the channel surface potentials[26]. Through this coupling, the model naturally captures pulse-width- and amplitude-dependent programming and multi-level conductance modulation.

Fig. 2a shows the simulated transfer characteristics for L=1 μm and W=1 μm (scaled as the basis for 32-nm node simulations) for both the programmed and erased device states. The resulting memory window, defined by the difference between the threshold voltages of these two states, is strongly influenced by the characteristics of the applied gate voltage waveform. The compact model captures this dependence in real time, enabling accurate prediction of threshold voltage modulation under varying programming conditions. Fig. 2b illustrates the sensitivity of the memory window to the pulse duration and the rise/fall times of the programming signal. Furthermore, we have validated the model against experimentally measured ULTRA**RAM** characteristics, as shown in Fig. 2c, demonstrating close agreement between simulation and measurement.

III. DNNS USING ULTRARAM SYNAPSE

The in-memory computing (IMC) architecture accelerates convolutional-neural-network (CNN) processing by executing matrix-vector multiplications directly within the memory crossbar array. The fundamental concept of analog IMC is to represent weights as conductance states within memory cells, mimicking synaptic behavior. In this work, we have utilized an ULTRA**RAM** memory device as a synapse, which enables the storage of multiple conductance states. First, we have employed experimentally demonstrated ULTRA**RAM** cells to evaluate the actual on-chip performance. Since the currently fabricated devices have relatively long channel lengths (∼10 μm) and no other emerging memory technologies are available at this scale, their performance has been compared against conventional SRAM-based synapses to provide a consistent estimation of performance metrics. Additionally, they exhibit limited conductance states (2-bit) suitable for verifying device physics but not for high-accuracy neuromorphic performance. Therefore, to project the technology's competitive potential, we have simulated scaled-down devices at the 32-nm node using a compact model calibrated against our experimental long-channel data, which matches the current state-of-the-art feature sizes of other emerging memory technologies.

A. Device–Circuit–System Co-Design Methodology

A device-to-system-level co-design approach is employed to simulate on-chip learning of a convolutional neural network (CNN) implemented using ULTRA**RAM**-based synaptic devices, as illustrated in Fig. 3. The simulation flow begins at the device level and progressively propagates through circuit, architectural, and system levels, enabling consistent cross-layer evaluation of learning accuracy, energy consumption, and latency.

At the device level, a physics-based compact model of the ULTRA**RAM** synapse is used, as summarized in the previous section[19]. The model captures resonant tunneling–assisted program and erase dynamics, time-dependent floating-gate charge accumulation, and the resulting conductance modulation. This enables realistic emulation of multi-level synaptic states, pulse-dependent weight updates, and intrinsic device variability during on-chip learning.

Using the proposed model, synaptic crossbar arrays of size 128×128 are implemented in a SPICE circuit simulator for each neural network layer. The output of each column of the crossbar array is connected to an output neuron, allowing direct evaluation of analog vector–matrix multiplication through Kirchhoff's current law. During each training iteration, circuit-level simulations generate column currents that correspond to the weighted sum of the inputs. These outputs are then transferred to a Python-based neural network engine, where forward propagation, error computation, and weight updates are performed. The updated synaptic weights are translated into ULTRA**RAM**-specific programming pulses and applied back to the circuit-level model, thereby closing the training loop.

At the system level, image classification is evaluated using the CIFAR-10 dataset with a VGG-8 neural network architecture[27,28]. This co-simulation approach ensures that learning dynamics are influenced by realistic device and circuit non-idealities rather than idealized weight updates, enabling faithful assessment of ULTRA**RAM**-based on-chip learning.

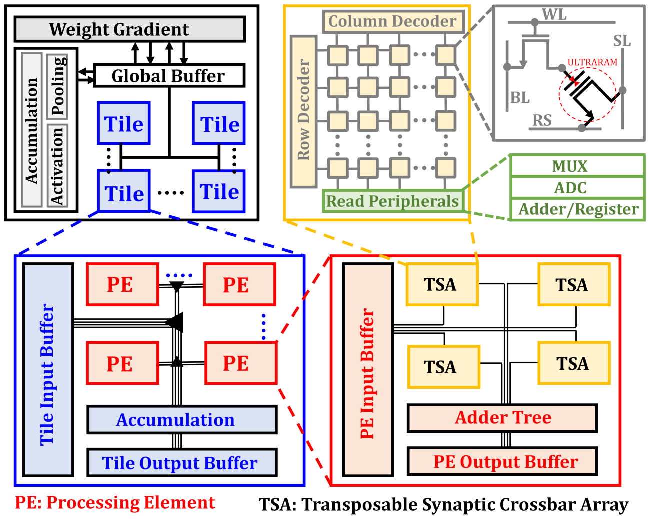


FIG. 3: Architecture-level representation of the ON-chip learning hardware.

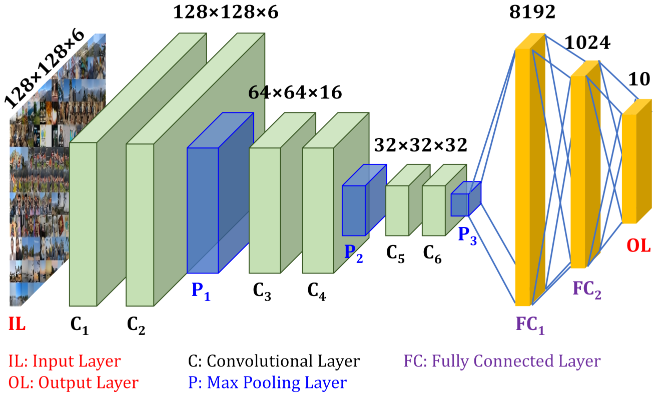
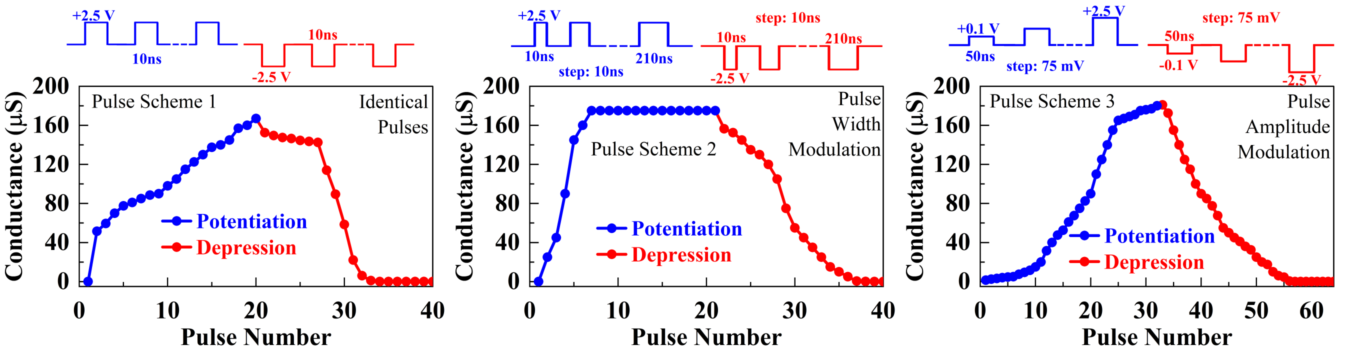


FIG. 4: Schematic of the VGG-8 model[27] used for image classification from the CIFAR-10 dataset[28].

B. Hardware Architecture for Neural Network Implementation

The hardware implementation for on-chip learning is shown in Fig. 3. The fundamental computing unit consists of ULTRA**RAM**-based crossbar arrays integrated with peripheral read/write circuits, analog-to-digital converters (ADCs), multiplexers, and adders, forming a transposable synaptic array (TSA). The pseudo-crossbar array consists of an access transistor paired with each memory cell, ensuring that only selected rows are programmed during row-wise weight updates and preventing unintended programming of unselected rows. ULTRA**RAM** synapses operate as three-terminal devices (with the back gate grounded) and require separate signals for word-line activation and read-select (RS) control. The RS signal enables retrieval of input vectors during read operations, as shown in Fig. 3. Multiple TSAs are interconnected using H-routing with embedded buffers to construct processing elements (PEs), which are then organized into tiles. Each tile includes dedicated units for weight-gradient computation,



(a) (b) (c)

FIG. 5: Simulated response of a scaled 32-nm ULTRA**RAM** cell to (a) identical pulses (same magnitude and pulse width), (b) variable pulse width for a fixed voltage magnitude, and (c) variable amplitude for a fixed pulse width. The number of accessible partial states is maximized when using a variable amplitude pulse scheme (∼ 32 states for LTP and LTD).

global buffering, accumulation, activation, and pooling, enabling parallel execution across neural network layers. Weight updates are performed sequentially in a row-by-row manner, while inference is executed in parallel by activating all columns simultaneously. Write and read lines regulate access transistors, enabling selective read and write operations for individual synaptic devices. To optimize energy and area efficiency, column multiplexing is employed, where one ADC is shared across eight columns. Along each column, output vectors are initially generated as analog partial current sums, which are subsequently digitized by the ADCs. Final accumulation of multi-state synaptic weights and input multiplications is carried out using shift-and-add digital processing modules.

C. Neural Network Architecture and Training Flow

The VGG-8 architecture is utilized for classifying 32×32 color images from the CIFAR-10 dataset, as illustrated in Fig. 4[27,28]. This network comprises six convolutional layers (C_1–C_6) for feature extraction, followed by two fully connected layers (FC_1 and FC_2) for image classification. Max-pooling layers with a 2×2 kernel are applied after each convolutional layer to downsample feature maps. During inference, input voltages corresponding to extracted image features are applied to the word lines of the ULTRA**RAM**-based crossbar arrays. The resulting bit-line currents represent the element-wise multiplication of input activations and synaptic conductance and are accumulated according to Kirchhoff's law. These outputs are then digitized and processed through activation circuits at each output node, enabling efficient in-memory matrix–vector multiplication.

For training, stochastic gradient descent is used to compute weight updates at each output node. The calculated weight changes are multiplied by the corresponding input activations using dedicated multiplier circuits. The resulting voltages serve as programming pulses for the ULTRA**RAM** synapses, adjusting their conductance states to reflect updated weight values.

IV. NON-IDEAL SYNAPTIC DEVICE PROPERTIES

The conductance of synaptic devices can be adjusted by applying positive or negative programming voltage pulses, corresponding to weight increment and decrement, respectively. Ideally, a synaptic device exhibits a linear weight update response to uniform programming voltage pulses. However, practical devices might deviate from this ideal behavior, displaying "non-ideal" characteristics such as nonlinear and fluctuating weight updates. This can restrict precision and lead to a finite ON/OFF ratio. We have analyzed the long-term potentiation (LTP) and long-term depression (LTD) behavior of ULTRA**RAM** devices under different pulse schemes. Fig. 5a shows the Scheme 1 with identical pulses. Each programming pulse has the same amplitude and duration for both potentiation and depression. In Scheme 2, the applied pulse width is varied gradually, keeping magnitude constant, to control the weight update, as shown in Fig. 5b. Lastly, in Scheme 3, we have applied a fixed time period pulse (50ns) width varying pulse magnitude from $\pm 0.1V$ to $\pm 2.5V$, as shown in Fig. 5c. The Scheme 3 shows the linear weight update in both potentiation and depression compared to other two schemes. In addition, it provides the maximum number of accessible partial states compared to the other schemes. The conductance change with a number of pulses (P) is fitted and nonlinearity in LTP and LTD are extracted by the method in the DNN+NeuroSim Framework[21] as follows:

$$G_{LTP} = B\left(1 - exp\left(-\frac{P}{\alpha_p}\right)\right) + G_{min} \tag{4}$$

$$G_{LTD} = -B\left(1 - exp\left(\frac{P - P_{max}}{\alpha_d}\right)\right) + G_{max} \tag{5}$$

$$B = (G_{max} - G_{min}) \Big/ \left(1 - exp\left(\frac{-P_{max}}{\alpha_{p,d}}\right)\right) \tag{6}$$

where, G_{LTP} and G_{LTD} are the conductance for LTP and LTD, respectively. G_{max}, G_{min} and P_{max} are the maximum conductance, minimum conductance and the maximum pulse number

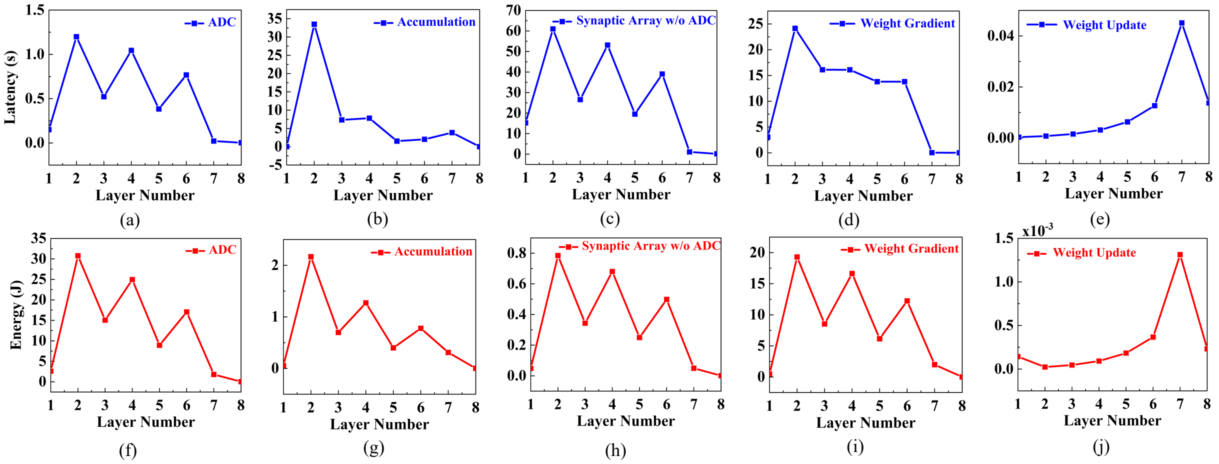


FIG. 6: (a)-(e) Peak latency and (f)-(j) energy across all the layers in VGG-8 for various CNN modules/operations (ADC, accumulation, synaptic array, weight gradient calculation, and weight update) in one epoch. The data shown is from the 256th epoch of 2-bit ULTRA**RAM**-based CIM architectures.

required to switch the device between the minimum and maximum conductance states, respectively. $\alpha_{p,d}$ is the parameter that controls the nonlinear behavior of weight update, and B is simply a function of $\alpha_{p,d}$ that fits the functions within the range of G_{max}, G_{min} and P_{max}. Scheme 3 exhibits the greatest number of states with symmetric response due to optimal sampling of charge storage in the FG through TBRT. Therefore, we have considered this scheme for on-chip training using ULTRA**RAM** cells.

V. PERFORMANCE OF CNN

The performance of CNNs was evaluated using experimentally demonstrated long-channel-length ULTRA**RAM** cells, and projected the performance with simulated devices at scaled technology nodes. A physics-based model has been used to investigate the experimental and theoretical response of ULTRA**RAM** cells for various pulse schemes. The model captures trapping and de-trapping in the floating gate of the ULTRA**RAM** devices through TBRT. The current density in the TBRT structure is calculated using a multi-barrier resonant tunneling current formulation. Further, the floating gate charge is used to determine the threshold voltage shift in the program and erase states. A detailed description of the model can be found in[19,20]. Then, a synaptic crossbar array of size 128×128 has been considered for simulations using the DNN+NeuroSIM simulator for each layer separately.

A. Long-channel Devices

We have considered two types of long-channel device for on-chip performance simulations: (1) ULTRA**RAM** cells fabricated on GaAs and Si substrates with 10μm of channel lengths[18,23]. These devices exhibit a limited current ratio, which restricts the number of achievable conductance states (2-bit), as shown in Fig. 2c. Nevertheless, appropriate device design and optimization can significantly improve their output characteristics upto 5-bit/cell with similar device dimensions[29], and discussed later in this section. (2) We have also considered these improved characteristics ULTRA**RAM** cells (5-bit) with similar device dimensions and used to predict the potential on-chip performance with optimized properties. This can serve as design guidelines for advancing present ULTRA**RAM** technology.

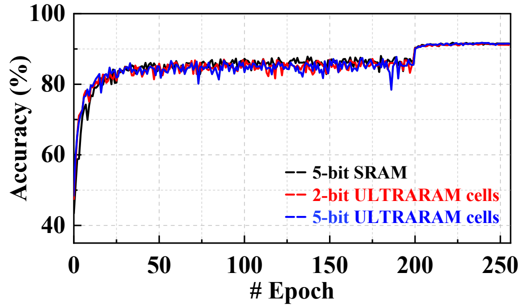


FIG. 7: Accuracy achieved for 5-bit SRAM, 2-bit experimentally demonstrated ULTRA**RAM**, and 5-bit simulated ULTRA**RAM** device precision in 256 epochs.

The full set of performance metrics is obtained over 256 epochs. Fig. 6 shows the latency and energy consumption for each layer of various CNN modules and operations. This includes the ADC, accumulation, synaptic array, weight gradient computation, and weight update. The final layer of the VGG-8 model has the smallest computational load because it maps the smallest feature vector to only 10 output classes. Additionally, it contains the fewest weights and requires the least MAC operations, leading to the lowest energy and latency. The overall energy and latency are primarily influenced by four key processes: feedforward, error computation, gradient computation, and weight update. Among these, weight gradient computation significantly impacts both energy and latency due to the frequent read and write operations required for activation functions and error processing.

To assess the influence of an ULTRA**RAM** synapse on

TABLE I: Benchmark results of CIM accelerators training on VGG-8 for CIFAR-10, based on SRAM and Long-channel ULTRA**RAM** synaptic cells with 256 epochs.

Technology Node	**130-nm**			
Device	**SRAM**	**ULTRARAM (GaAs Subs.)**[18]	**ULTRARAM (Si Subs.)**[23]	**ULTRARAM (Optimized)***
# Conductance States	32	4	4	32
Cell Precision	1-bit	2-bit	2-bit	5-bit
R_{ON} [Ω]	–	$0.6K$	$0.33K$	$5K$
ON/OFF Ratio	–	2	2	10
C2C Variation	–	<0.5%	<0.5%	3%
Write Pulse Voltage [V]	–	±2.5	±2.5	±2.5
Write Pulse Width	–	500 μs	10 ms	100 ns
Area [mm^2]	6295.3	3491	3576	1862
Training Accuracy	91.7	91.52	91.68	91.69
Training Latency (s) / Epoch	453.2	490.4	588	362.12
Training Dynamic Energy (J) / Epoch	358.4	235.43	267	173.6
Training Throughput (TOPS)	0.406	0.376	0.31	0.50
Training Energy Efficiency (TOPS/W)	0.508	0.781	0.68	1.06

*Projected performance from long channel devices with optimized characteristics.

a CNN's performance, the proposed 2-bit and 5-bit ULTRA**RAM**-based CNNs were evaluated in comparison to a 5-bit SRAM-based CNN using the same simulation framework. Fig. 7 shows the relationship between the number of training epochs and the accuracy of 5-bit SRAM and two different ULTRA**RAM** cells implemented with 2-bit and 5-bit weight precision. It is observed that the ULTRA**RAM**-based neural network demonstrates accuracy comparable to that of a 5-bit SRAM-based design. However, the 2-bit ULTRA**RAM**-based CNN exhibits superior efficiency, being 1.8× more area-efficient and 1.52× more energy-efficient. However, it loses in terms of latency and can be seen in Table I. For a fair comparison, we have compared 5-bit SRAM with a 5-bit ULTRA**RAM**-based CNN. The 5-bit ULTRA**RAM** cells have been simulated in TCAD with the channel of $1\mu m$ and TBRT quantum well thicknesses of $3nm$ and $2.4nm$. We have observed the $100ns$ switching time during the program/erase operations with the input pulse of $\pm 2.5V$. This results in improvement in area, energy, and latency by 3.38×, 2.06×, and 1.25×, respectively, compared to 5-bit SRAM-based CNN without affecting the accuracy and can be seen in Fig. 7.

Finally, we have evaluated the performance of CIM accelerators for VGG-8 training on the CIFAR-10 dataset[27,28], utilizing ULTRA**RAM** and SRAM-based accelerators. Due to the longer channel lengths ($>10\mu m$) of experimentally demonstrated ULTRA**RAM** cells, we have assumed 130 nm technology node for evaluating the on-chip performance. Table I shows the benchmark results of CIM accelerators based on SRAM and ULTRA**RAM** synaptic cells with 256 epochs. The on-chip 5-bit SRAM-based CMOS implementation provides the same training accuracy but requires a significantly larger chip area overhead relative to 2-bit ULTRA**RAM** non-volatile memory cells. Additionally, the 2-bit ULTRA**RAM** synapses exhibit a comparable energy, latency and TOPS advantage compared to 5-bit SRAM-based synapses. These performance parameters can be further improved by using a optimized 5-bit ULTRA**RAM**-based synapses, as projected in Table I.

B. Projection with Scaled Devices

While fabrication of sub-micron devices is ongoing, we have simulated the ULTRA**RAM** cells with scaled-down channel lengths ($\sim 100\ nm$) considering the same TBRT stack replacing the gate oxide. Now, we have compared this with other analog emerging memory devices at 32-nm technology nodes.

Fig. 8 shows the latency and energy consumption for each layer of various CNN modules and operations considering the 5-bit ULTRA**RAM**-based synapse. This shows that the latency and energy consumption can be significantly reduced with the scaled ULTRA**RAM** cells as compared to experimentally demonstrated cells [Fig. 6]. In addition, the training accuracy is comparable to the existing ULTRA**RAM** cells with 3% of cycle-to-cycle (C2C) variations, as shown in Fig. 9. We have used the pulse Scheme 3 (pulse amplitude modulation) to plot the conductance change with the number of pulses (P) and non-linearity in LTP and LTD using the equations (4) and (5), as shown in Fig. 10. The 5-bit ULTRA**RAM**-based CNN exhibits better efficiency, being 1.36× more area-efficient, 1.1× more energy-efficient, and 1.87× faster in terms of latency compared to 32-nm node SRAM-based CNN.

Finally, we have benchmarked the performance of CIM accelerators utilizing various analog synaptic devices, including memristor[30], RRAM[16,31], EpiRAM[32], and FeFET[14], with ULTRA**RAM**-based synapse at 32-nm technology node, as

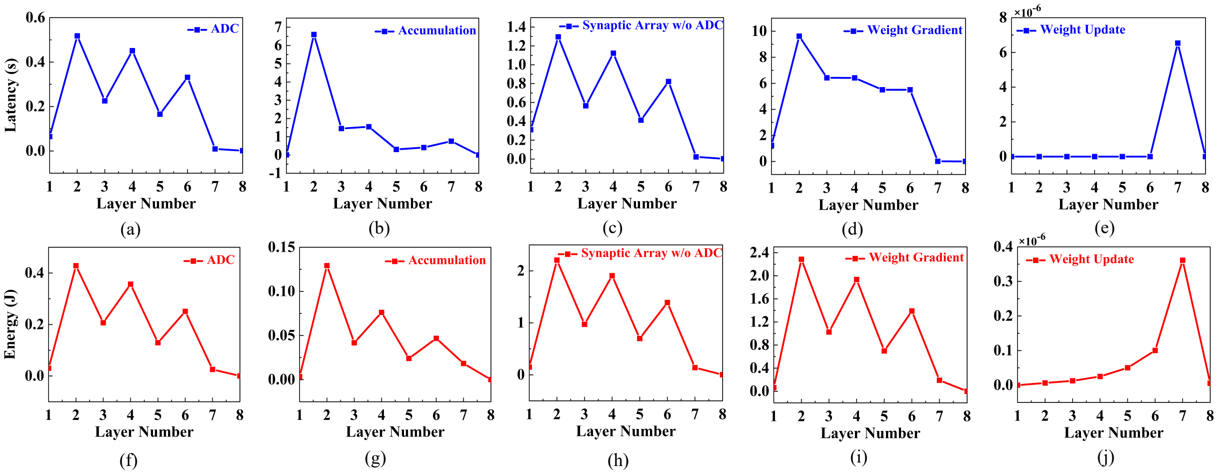


FIG. 8: (a)-(e) Peak latency and (f)-(j) energy across all the layers in VGG-8 for various CNN modules/operations (ADC, accumulation, synaptic array, weight gradient calculation, and weight update) in one epoch. The data shown is from the 256th epoch of simulated 5-bit ULTRA**RAM**-based CIM architecture.

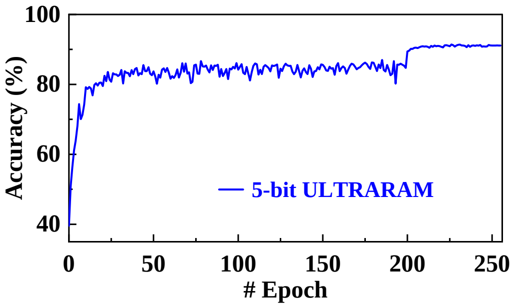


FIG. 9: Accuracy achieved in 256 epochs of 5-bit ULTRA**RAM**-based CIM architecture at 32-nm technology node.

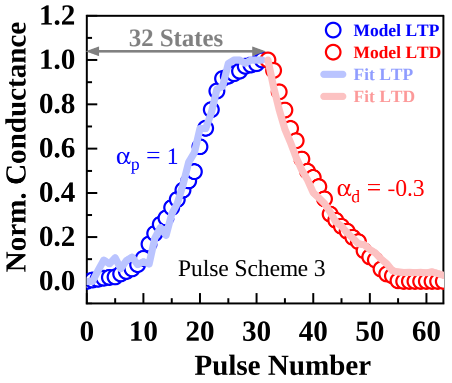


FIG. 10: Normalized simulated response of a 32-nm node ULTRA**RAM** cell using pulse Scheme 3 (varying magnitude with a fixed pulse width). The corresponding non-linearity ($\alpha_{p/d}$) has been extracted using the equations (4) and (5).

shown in Table II. It is observed that the ULTRA**RAM**-based synapse can provide better performance in terms of throughput, area, latency, and energy compared to SRAM. This is attributed to the underlying switching physics of the synaptic devices. ULTRA**RAM** achieves low write energy and fast programming by leveraging resonant tunneling–assisted charge transport across a triple-barrier structure, enabling rapid transitions between high- and low-resistance states with relatively low programming voltages (±2.5 V)[23]. In contrast, PCM relies on thermally driven phase transitions that incur substantial Joule heating, whereas RRAM relies on ionic filament formation and rupture, both of which result in higher write energy and additional latency overhead[13,16]. FeFETs, being voltage-driven devices with comparable pulse amplitudes and widths, exhibit system-level energy and latency performance similar to ULTRA**RAM**. Furthermore, the intrinsically fast carrier tunneling dynamics in ULTRA**RAM** eliminate the need for iterative write-verify operations or thermal stabilization, allowing programming times on the order of tens of nanoseconds. This framework integrates device-level, circuit-level, and architectural non-idealities into the simulations and allows us to capture realistic on-chip training behavior, including IR drops, write noise, device variations, and peripheral circuit overheads. Therefore, the reported performance metrics are based on hardware-aware simulations.

ULTRA**RAM** memory shows promise as a synaptic cell for DNN acceleration. Based on the hardware performance results presented in Tables I and II, the following observations can be made: (i) Optimizing on-state resistance (R_{ON}) is critical for minimizing voltage drops; however, scaling transistors in 1T1R architectures or peripheral multiplexers increases area overhead and parasitic capacitance, adversely impacting latency and throughput. (ii) Write pulse durations below a microsecond remain unaffected due to batch-wise amortization. (iii) Maintaining cycle-to-cycle variation below 1% is essential to ensure stable in-situ training, as higher variations can disrupt model convergence. (iv) While SRAM-based architectures encounter leakage and area constraints at larger tech-

TABLE II: Benchmark results of CIM accelerators training on VGG-8 for CIFAR-10, based on SRAM, reported analog synaptic devices, and ULTRA**RAM** synaptic cells with 256 epochs.

Technology Node	**32-nm**						
Device	**SRAM**	**Memristor** [30]	**RRAM (PCMO)** [31]	**RRAM (AlO_x/HfO_2)** [16]	**EpiRAM** [32]	**FeFET** [14]	**ULTRARAM*** **(This Work)**
# Conductance States	–	97	50	40	64	32	32
Cell Precision	1	6	5	5	6	5	5
R_{ON} [Ω]	–	26*M*	23*M*	16.9*K*	81*K*	240*K*	5*K*
ON/OFF Ratio	–	12.5	6.84	4.43	50.2	10	10
C2C Variation (%)	–	3.5	<1	5	2	<0.5	3
Write Pulse Voltage [V]	–	±3	±2	±1	±5	±4	±2.5
Write Pulse Width	–	300 μs	1 *ms*	100 μs	5 μs	50 *ns*	50 *ns*
Area [mm^2]	138.95	48.29	48.29	49.88	48.59	95.21	101.48
Training Accuracy (%)	91	49	56	37	85	91.12	91.28
Training Latency (s) / Epoch	235.75	1241.63	5795.79	611	193.94	121.66	125.9
Training Dynamic Energy (J) / Epoch	95.37	92.12	92.15	93.13	92.28	87.18	86.68
Training Throughput (TOPS)	0.78	0.14	0.003	0.30	0.95	1.51	1.46
Training Energy Efficiency (TOPS/W)	1.94	2	2	1.98	2	2.11	2.12

*Projected performance with 32-nm technology node scaled device parameters simulated with model.

nology nodes, parallel-read SRAM designs at advanced nodes offer superior energy efficiency and throughput.

VI. CONCLUSIONS

In this work, we have presented on-chip training and inference of a neural network using ULTRA**RAM** memory device-based synaptic arrays. The longer channel 2-bit ULTRA**RAM**-based CNN exhibits superior efficiency, being 1.8× more area-efficient and 1.52× more energy-efficient. Additionally, the performance projection has been demonstrated with the simulated ULTRA**RAM** cells scaled down to advanced technology nodes (32-nm). This results superior performance than SRAM- and several emerging memory technologies-based CNN implementations, while maintaining performance levels comparable to FeFET-based designs with respect to critical system metrics such as area, latency, energy consumption, and throughput. ULTRA**RAM** shows considerable promise for enabling efficient synaptic operations in DNN accelerators.

ACKNOWLEDGMENTS

This work was supported in parts by the Quinas Technology Limited, Lancaster, United Kingdom; Indian Institute of Technology Roorkee, India, and in part by the Prime Minister's Research Fellowship, Ministry of Education, Government of India under Grant PM-31-22-773-414.

AUTHOR DECLARATIONS

Conflict of Interest

M. Hayne and P. D. Hodgson are (part-time) employees and co-founding shareholders of Quinas Technology. M. Hayne is (co-)inventor of related pending and granted patents and P. D. Hodgson is co-inventor of related pending patents.

DATA AVAILABILITY

The data that support the findings of this study are available within the article.

REFERENCES

[1] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural networks*, vol. 61, pp. 85–117, 2015.

[2] N. Rusk, "Deep learning," *Nature Methods*, vol. 13, no. 1, pp. 35–35, 2016.

[3] H.-S. P. Wong and S. Salahuddin, "Memory leads the way to better computing," *Nature nanotechnology*, vol. 10, no. 3, pp. 191–194, 2015.

[4] C.-J. Jhang, C.-X. Xue, J.-M. Hung, F.-C. Chang, and M.-F. Chang, "Challenges and trends of sram-based computing-in-memory for ai edge devices," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 68, no. 5, pp. 1773–1786, 2021.

[5] K. Yu, S. Kim, and J. R. Choi, "Trends and challenges in computing-in-memory for neural network model: A review from device design to application-side optimization," *IEEE Access*, 2024.

[6] S. Mittal, G. Verma, B. Kaushik, and F. A. Khanday, "A survey of sram-based in-memory computing techniques and applications," *Journal of Systems Architecture*, vol. 119, p. 102276, 2021.

[7] F. Gao, G. Tziantzioulis, and D. Wentzlaff, "Computedram: In-memory compute using off-the-shelf drams," in *Proceedings of the 52nd annual IEEE/ACM international symposium on microarchitecture*, 2019, pp. 100–113.

[8] S. Khoram, Y. Zha, J. Zhang, and J. Li, "Challenges and opportunities: From near-memory computing to in-memory computing," in *Proceedings of the 2017 ACM on International Symposium on Physical Design*, 2017, pp. 43–46.

[9] S. Kim and H.-J. Yoo, "An overview of computing-in-memory circuits with dram and nvm," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 71, no. 3, pp. 1626–1631, 2024.

[10] S. Dutta, H. Ye, W. Chakraborty, Y.-C. Luo, M. San Jose, B. Grisafe, A. Khanna, I. Lightcap, S. Shinde, S. Yu *et al.*, "Monolithic 3d integration

of high endurance multi-bit ferroelectric fet for accelerating compute-in-memory," in *2020 IEEE International Electron Devices Meeting (IEDM)*. IEEE, 2020, pp. 36–4.

[11] S. Yin, Y. Kim, X. Han, H. Barnaby, S. Yu, Y. Luo, W. He, X. Sun, J.-J. Kim, and J.-s. Seo, "Monolithically integrated rram-and cmos-based in-memory computing optimizations for efficient deep learning," *IEEE Micro*, vol. 39, no. 6, pp. 54–63, 2019.

[12] G. Pedretti and D. Ielmini, "In-memory computing with resistive memory circuits: Status and outlook," *Electronics*, vol. 10, no. 9, p. 1063, 2021.

[13] Q. Wang, G. Niu, W. Ren, R. Wang, X. Chen, X. Li, Z.-G. Ye, Y.-H. Xie, S. Song, and Z. Song, "Phase change random access memory for neuro-inspired computing," *Advanced Electronic Materials*, vol. 7, no. 6, p. 2001241, 2021.

[14] M. Jerry, P.-Y. Chen, J. Zhang, P. Sharma, K. Ni, S. Yu, and S. Datta, "Ferroelectric fet analog synapse for acceleration of deep neural network training," in *2017 IEEE International Electron Devices Meeting (IEDM)*, 2017, pp. 6.2.1–6.2.4.

[15] J. Yoo, H. Song, H. Lee, S. Lim, S. Kim, K. Heo, and H. Bae, "Recent research for hzo-based ferroelectric memory towards in-memory computing applications," *Electronics*, vol. 12, no. 10, p. 2297, 2023.

[16] J. Woo, K. Moon, J. Song, S. Lee, M. Kwak, J. Park, and H. Hwang, "Improved synaptic behavior under identical pulses using alox/hfo2 bilayer rram array for neuromorphic systems," *IEEE Electron Device Letters*, vol. 37, no. 8, pp. 994–997, 2016.

[17] P.-Y. Chen, X. Peng, and S. Yu, "Neurosim+: An integrated device-to-algorithm framework for benchmarking synaptic devices and array architectures," in *2017 IEEE International Electron Devices Meeting (IEDM)*. IEEE, 2017, pp. 6–1.

[18] D. Lane, P. Hodgson, R. Potter, R. Beanland, and M. Hayne, "ULTRARAM: toward the development of a iii–v semiconductor, nonvolatile, random access memory," *IEEE Transactions on Electron Devices*, vol. 68, no. 5, pp. 2271–2274, 2021.

[19] A. Kumar, M. Ehteshamuddin, A. Bulusu, S. Mehrotra, and A. Dasgupta, "A physics-based compact model for ultraram memory device," in *2024 8th IEEE Electron Devices Technology and Manufacturing Conference (EDTM)*, 2024, pp. 1–3, doi: 10.1109/EDTM58488.2024.10512293.

[20] A. Kumar and A. Dasgupta, "Compact modeling of compound semiconductor memory ultraram: A universal memory device," in *2024 Device Research Conference (DRC)*, 2024, pp. 1–2, doi: 10.1109/DRC61706.2024.10605295.

[21] X. Peng, S. Huang, H. Jiang, A. Lu, and S. Yu, "Dnn+neurosim v2.0: An end-to-end benchmarking framework for compute-in-memory accelerators for on-chip training," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 40, no. 11, pp. 2306–2319, 2021.

[22] Experiment limited. Zero degradation observed after 10^7 program/erase cycles.

[23] P. D. Hodgson, D. Lane, P. J. Carrington, E. Delli, R. Beanland, and M. Hayne, "ULTRARAM: A low-energy, high-endurance, compound-semiconductor memory on silicon," *Advanced Electronic Materials*, vol. 8, no. 4, p. 2101103, 2022.

[24] D. Lane and M. Hayne, "Simulations of resonant tunnelling through inas/alsb heterostructures for ULTRARAM memory," *Journal of Physics D: Applied Physics*, vol. 54, no. 35, p. 355104, 2021.

[25] D. Lane, P. Hodgson, R. Potter, and M. Hayne, "Demonstration of a fast, low-voltage, III-V semiconductor, non-volatile memory," in *2021 5th IEEE Electron Devices Technology & Manufacturing Conference (EDTM)*. IEEE, 2021, pp. 1–3.

[26] S. Khandelwal, Y. S. Chauhan, and T. A. Fjeldly, "Analytical modeling of surface-potential and intrinsic charges in algan/gan hemt devices," *IEEE Transactions on Electron Devices*, vol. 59, no. 10, pp. 2856–2860, 2012.

[27] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015.

[28] M. A. Rasslan, "Alexnet, vgg16, and vgg8 on cifar-10," Kaggle Notebook, 2025, https://www.kaggle.com/code/mennaalaarasslan/alexnet-vgg16-and-vgg8-on-cifar-10.

[29] A. Kumar, M. Dar, P. Hodgson, D. Lane, P. Carrington, E. Delli, R. Beanland, S. Mehrotra, M. Hayne, and A. Dasgupta, "Physics, modeling, and benchmarking of ultraram: A compound semiconductor-based memory device," *Journal of Applied Physics*, vol. 138, no. 9, 2025, doi: 10.1063/5.0269780.

[30] S. H. Jo, T. Chang, I. Ebong, B. B. Bhadviya, P. Mazumder, and W. Lu, "Nanoscale memristor device as synapse in neuromorphic systems," *Nano letters*, vol. 10, no. 4, pp. 1297–1301, 2010.

[31] S. Park, A. Sheri, J. Kim, J. Noh, J. Jang, M. Jeon, B. Lee, B. Lee, B. Lee, and H.-J. Hwang, "Neuromorphic speech systems using advanced reram-based synapse," in *2013 IEEE International Electron Devices Meeting*. IEEE, 2013, pp. 25–6.

[32] S. Choi, S. H. Tan, Z. Li, Y. Kim, C. Choi, P.-Y. Chen, H. Yeon, S. Yu, and J. Kim, "Sige epitaxial memory for neuromorphic computing with reproducible high performance based on engineered dislocations," *Nature materials*, vol. 17, no. 4, pp. 335–340, 2018.