RRAM-enabled AI Accelerator Architecture

Xinxin Wang[1,2], Yuting Wu[1,2], and Wei D. Lu[1]*
[1]Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, Michigan 48109, USA
[2]These authors contributed equally to this work.
*Email: wluee@umich.edu

***Abstract*—**Resistive random-access memory (RRAM) offers high-density non-volatile storage and potential for efficient in-memory computing (IMC). RRAM-enabled accelerators can solve the von Neumann bottleneck and meet the ever-growing computing needs of applications such as Artificial Intelligence (AI). In this paper, we discuss progress and challenges in RRAM-based accelerators for AI inference, training and arithmetic applications. Architecture strategies to accommodate large neural networks will be introduced. Alternative computing systems enabled by the devices' internal dynamics will also be presented.

I. Introduction

The exponential growth of AI workloads calls for highly efficient hardware, a task that becomes increasingly challenging to meet through device scaling alone. IMC accelerators based on emerging memory devices such as RRAM offer high compute density, throughput and energy efficiency, and have generated broad interest for these tasks.

In general, vector-matrix multiplication (VMM), the core operation in many AI workloads, can be performed directly in RRAM arrays in a single step. Here we will review recent implementations of RRAM accelerator prototypes for AI inference and training applications, along with approaches to perform efficient logic and high precision arithmetic operations. We will discuss challenges including ADC overhead and device nonidealities and introduce a modular architecture design to accommodate large AI models. Finally, approaches to perform computing beyond multiply-accumulate (MAC) by taking advantage of the internal device dynamics, including synaptic plasticity learning, reservoir computing (RC) and stochastic computing (SC), will be presented.

II. Resistive Switching Devices

RRAM is a type of resistive switching device that exhibits reversible resistance changes through physical reconfiguration of the material composition. There are two main types of RRAM: one based on cation redistribution, called electrochemical metallization memory (ECM), also known as conductive bridge random access memory (CBRAM); and one based on anion redistribution, known as valency change memory (VCM) or oxide-RRAM, as shown in Fig. 1(a).

In both types, the devices operate based on the formation and rupture of a conductive filament, through the oxidation, migration and reduction of metal ions (e.g., Ag or Au) or oxygen vacancies, respectively. Examples of filaments formed after the SET progress are shown in Fig. 2 [1, 2]. The filament growth processes are in turn determined by internal thermodynamic and kinetic factors including ionization energy, barrier height for ion migration, hopping distance, and redox rates, leading to different filament growth modes [1, 3].

By self-consistently solving the electronic, ionic and thermal transport equations, the dynamic resistive switching processes can be accurately modelled (Fig. 3) [4], providing insights into the internal physical mechanisms for continued device optimizations and circuit simulations.

III. RRAM-Enabled AI Accelerators

Through Ohm's law and Kirchhoff's current law, an RRAM array can be used to directly perform MAC, particularly in its matrix form - VMM, in a single time step, as shown in Fig. 4 [5]. Additionally, by reversing the input and output, the same array can be used to perform VMM using the output activation and the transposed weight matrix to produce the reconstruction of the input. This operation can be used to calculate error during backpropagation for network training, or to implement lateral neuron inhibition in algorithms such as sparse coding [6].

A single layer perceptron was first implemented for 3×3 pixel bitmap classification using a 12×12 Al_2O_3/TiO_{2-x} RRAM array [7]. Implementations using passive crossbar arrays have been challenging due to the sneak current issue which affects the current output during MAC and the programming voltage delivery during weight update. Recent progress with passive RRAM arrays includes a 32x32 WO_x array for sparse coding [6], and a fully integrated, reprogrammable chip with a passive RRAM crossbar array directly fabricated on top of CMOS circuits for different tasks such as bi-layer networks and sparse coding, as shown in Fig. 6(e) [8].

1T1R arrays with integrated transistors as selectors allow larger arrays to be fabricated and have recently been offered through major foundries. Several groups have developed MAC circuits based on 1T1R RRAM arrays. A 128×64 HfO_2 1T1R array with high device yield (99.8%) and high precision (6-bit) has been used for image processing/classification [9] and reinforcement learning [10], offering 1.64 TOPS and ~119.7 effective TOPS/W. A system consisting of 8 TaO_x/HfO_x RRAM chips with 128×16 1T1R cells per chip has been developed to support both convolutional and fully-connected operations across multiple chips [11]. An effective hybrid training method was proposed to accommodate device variations and near software-accuracy (1.49% drop) for CIFAR-10 classification. A 128 TOPS/W neurosynaptic core with 64k RRAM and 256 integrate-and-fire neurons has been implemented to support dynamically reconfigurable dataflow, transpose VMM and probabilistic sampling [13]. An IMC module integrating binary 1T1R RRAM cells with low power current mode readout circuits has been designed to achieve

This work was supported in part by SRC and DARPA through the Applications Driving Architectures (ADA) Research Center, and by NSF through Award CCF-1900675.

2021 IEEE International Electron Devices Meeting (IEDM) | 978-1-6654-2572-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/IEDM19574.2021.9720543

16.95 TOPS/W and 98.8% accuracy for MNIST inference [14]. A 2Mb integrated RRAM based IMC macro was also recently developed with TSMC 22nm CMOS technology [15]. Binary neural networks (BNNs) [16], where weights and activations are quantized to 1-bit precision, offer another promising approach to use mature binary RRAM devices and low precision ADCs (or multi-level sense amplifiers). A BNN system using 128×64 binary RRAM devices has demonstrated accuracy of 98.5% for MNIST and 83.5% for CIFAR-10 datasets, achieving 158 GOPS and 24 TOPS/W [17].

IV. In-memory logic and arithmetic operations

Low power, highly parallel and reconfigurable bitwise logic and arithmetic computing schemes have been studied on the basis of voltage-divider effects in RRAM arrays. Fig. 7 shows an example of implementing wired-NOR logic. The NOR gates can be further used as the foundation to build full adders and other logic and arithmatic operations [18].

A scheme to achieve high-precision computing, beyond the native precision offered by the device, was proposed in [19]. An example of a high-precision partial differential equation (PDE) solver is shown in Fig. 8(a). The sparse coefficient matrix (Fig. 8(b)) in PDE can be divided into slices and mapped onto RRAM crossbar arrays (Fig. 8(c)). The effective precision can be extended through ADC quantization at each crossbar and shifting/add operations, as shown in Fig. 8(d). This system has been used to solve static and time-evolving problems, such as Poisson's equations, damped 2D wave equations and argon plasma reactor, with precision extending up to double-precision (64-bits).

V. Internal dynamics-based computing

Internal dynamics of RRAM devices such as temporal ionic drift/diffusion processes and switching nonlinearity and stochasticity, can enable highly efficient computing schemes. The ion drift and Joule heating, driven by the applied field, and spontaneous ion diffusion and heat dissipation offer an internal Ca^{2+}-like timing mechanism (Fig. 9), and can be used to implement biorealistic synaptic plasticity learning including long term plasticity, short term plasticity and spike timing-dependent plasticity (STDP). Volatile RRAM devices with inherent short-term memory effect and nonlinear dynamics can be utilized to implement RC systems [20, 21], in which a dynamic reservoir can nonlinearly project temporal inputs onto a high-dimensional feature space to make them linearly separable for subsequent processing (Fig. 10). An RC system designed to perform handwritten digit classification task used only 88 WO_x-based RRAM devices (Fig. 11) [21] and subsequent studies on these RC systems have achieved tasks such as long-term time-series forecasting. The stochastic switching behavior of $Cu/Al_2O_3/Pd$ CBRAM devices was used to implement SC systems and emulate simulated annealing of a spin-glass [22], as shown in Fig. 12.

VI. Challenges and Perspects

Several challenges remain that may impede further implementation of RRAM-based AI accelerators. First, high-precision ADC-based readout circuits are a significant overhead. ADCs with resolutions higher than 8-bit will dominate the power and area of the MAC system, while low-precision ADCs may degrade the accuracy. Second, performance can suffer from device non-idealities including c-c variations and finite on/off ratio. Third, the nonlinear and asymmetric conductance update observed in RRAM devices can severely degrade training accuracy.

A solution to the first problem is to introduce several ranges in the ADC design with low area and power overhead [23]. An alternate method is the use of BNNs, where readout can be greatly simplified [13]. The device non-ideality issues for the second and third problems can be mitigated through mixed-precision training [24], and by using differential RRAM pairs (Fig. 13) which improves the operation margin and alleviates the asymmetric weight update problem [25].

The array size used for IMC is typically limited to <1Mb to minimize parasitic capacitance and parasitic resistance effects. To accommodate state-of-the-art deep neural networks (DNNs) which usually contain millions of parameters, modular systems that utilize many arrays integrated together through digital interface have been proposed [23, 26, 27, 30-32]. For example, PRIME [30] and ISAAC [31] attempted to address the ADC overhead by splitting the weights onto multiple devices and utilizing bit-serial approach to feed in the input activations. Pipelayer [32] considered the inter- and intra- layer parallelism, and replaced ADCs with simple spike drivers. The end-to-end performance of typical DNN models and the impact of device and circuit nonidealities have been analyzed in [23, 27]. One example of mapping DNN models onto the modular architecture is shown in Fig. 14(a). Performance degradation induced by ADC quantization of the partial sums (Psum) (Fig. 14(b)) can be minimized through multi-range quantization (Fig. 14(c)) and architecture-aware training (Fig. 14(e)), which has also been shown to address other device non-idealities such as finite on-off ratio and device variability (Fig. 14(d)).

By using several types of IMC modules to minimize ADC overhead and balance the latency of different layers in typical DNNs (Fig. 15(a)), TAICHI [33] was able to map several classes of DNNs with high throughput and energy efficiency using a single design. A hierarchical mesh network-on-chip (HM-NoC) was used for data routing between modules. Power and area breakdown of the design at different technology nodes are summarized in Fig. 15(b), where the 8-bit ADC was still found to be a dominating factor for area and power. By analyzing the compute-weight ratio, models exceeding the chip storage capacity can be split carefully to allow the compute-bound layers to be mapped onto the IMC and memory-bound layers onto a global co-processor to make the IMC system "future-proof", as shown in Fig. 15(c).

VII. Conclusion

Significant progress has been made in RRAM-enabled accelerators, both at the module level and at the system architecture level. With the recent commercialization of RRAM and continued improvements in device, architecture and algorithm, such IMC systems offer great potential to lead to the implementation of high performance, energy efficient and scalable hardware systems for AI and other data-intensive applications in the near future.

References

[1] Y. Yang *et al.*, *Nat. Commun.*, 2012. [2] M. J. Lee, *et al.*, *Nat. Mater.* 2011. [3] Y. Yang *et al.*, *Nat. Commun.,* 2014. [4] S. H. Lee *et al., ACS Appl. Electron. Mater.* , 2020. [5] M. Hu *et al.*, *DAC*, 2016. [6] P. M. Sheridan *et al.*, *Nat. Nanotechnol.,* 2017. [7] M. Prezioso *et al.*, *Nature,* 2015. [8] F. Cai *et al. Nat. Electron*. 2019. [9] C. Li *et al.*, *Nat. Electron.,* 2018. [10] Z. Wang *et al.*, *Nat. Electron.,* 2019. [11] P. Yao *et al.*, *Nature,* 2020. [12] Q. Liu, *et al., ISSCC. IEEE*, 2020. [13] W. Wan *et al., ISSCC. IEEE*, 2020. [14] W. H. Chen *et al., Nat. Electron.*, 2019. [15] C. X. Xue *et al., ISSCC. IEEE*, 2020. [16] M. Courbariaux *et al.*, *arXiv preprint arXiv:1602.02830,* 2016. [17] S. Yin *et al.*, *IEEE TED,* 2020. [18] B. Chen *et al.*, *IEDM,* 2015. [19] M. A. Zidan *et al.*, *Nat. Electron,* 2018. [20] C. Du *et al.*, *Nat. Commun.,* 2017. [21] X. Zhu *et al.*, *Nat Commun, 2020.* [22] J. H. Shin *et al.*, *IEDM,* 2018. [23] X. Wang *et al.*, *IEDM,* 2019. [24] M. Le Gallo *et al., Nat. Electron,* 2018. [25] T. Hirtzlin *et al.*, *IEDM,* 2019. [26] M. A. Zidan *et al.*, *IEEE T Multi-Scale C,* 2018. [27] X. Peng *et al, IEDM,* 2019. [28] Y. Liao *et al., IEEE TED,* 2020. [29] Q. Wang *et al., ISCAS,* 2021. [30] P. Chi *et al., SIGARCH,* 2016. [31] A. Shafiee *et al., SIGARCH,* 2016. [32] L. Song *et al., HPCA,* 2017. [33] X. Wang, *et al. IEEE TCAS-II,* 2021.

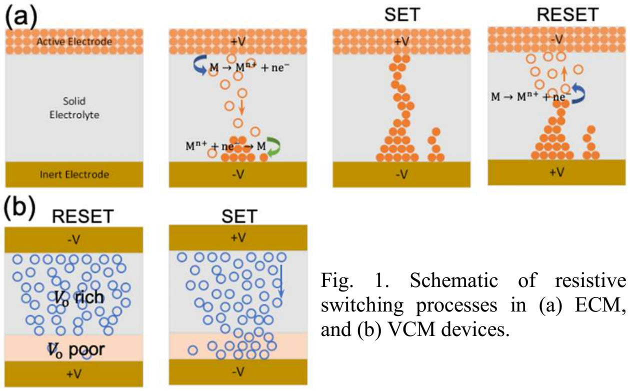

Fig. 1. Schematic of resistive switching processes in (a) ECM, and (b) VCM devices.

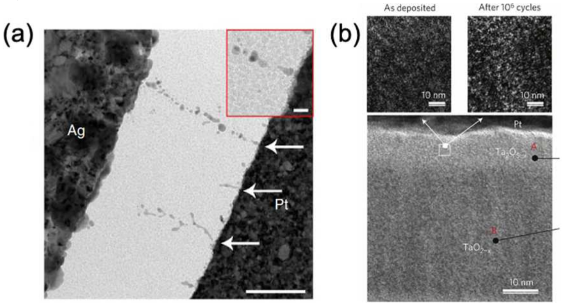

Fig. 2. (a) TEM image showing the formation of a Ag conductive filament after the forming process. Adapted from [1]. (b) High-resolution TEM image of a Pt/Ta_2O_{5-x}/TaO_{2-x}/Pt structure (bottom), showing the Ta_2O_{5-x} layer before cycling (top left) and after 10^6 cycles (top right) that exhibits Ta-rich clusters. Adapted from [2].

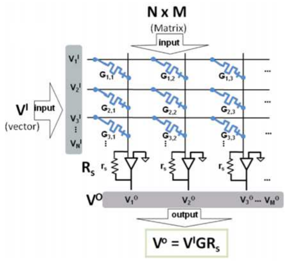

Fig. 4. Schematic of analog VMM implemented on RRAM-crossbar array. Adapted from [5].

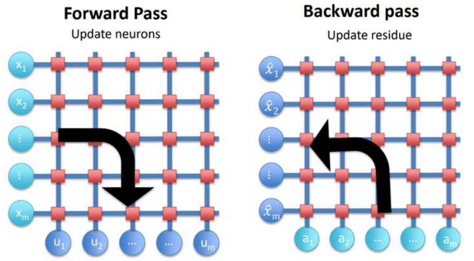

Fig. 5. Bi-directional RRAM operations. The forward pass performs VMM using the input vector and the stored weight matrix, and the backward pass performs VMM using the neuron activity vector with the transpose of the weight matrix. Adapted from [6].

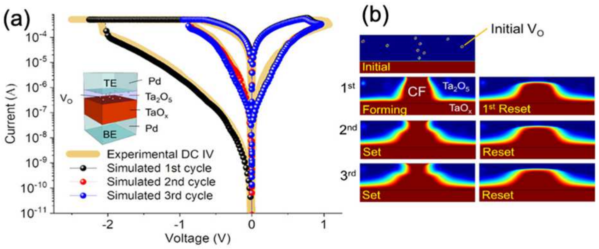

Fig. 3. (a) Measured and simulated I-V characteristics of tantalum oxide RRAM. Inset shows the device structure. (b) 2D maps of Vo concentration (n_D) obtained in the model, for the forming, first reset and subsequent switching cycles. Adapted from [4].

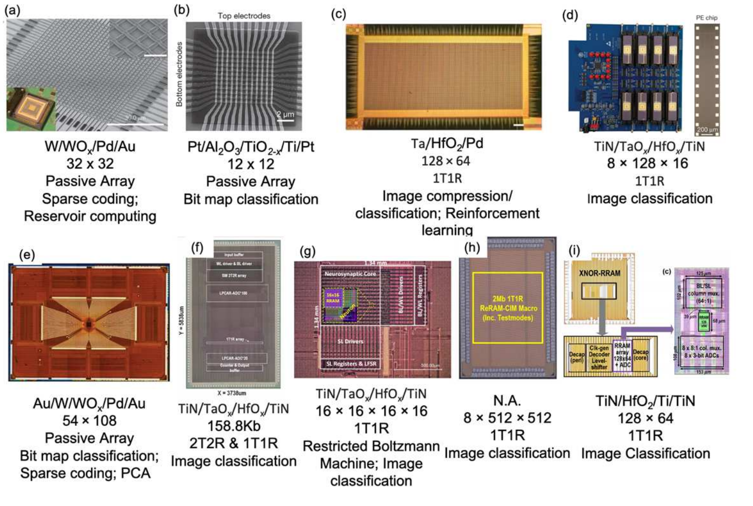

Fig. 6. RRAM-based AI accelerator prototypes. (a) Adapted from [6]. (b) Adapted from [7]. (c) Adapted from [9]. (d) Adapted from [11]. (e) Adapted from [8]. (f) Adapted from [12]. (g) Adapted from [13]. (h) Adapted from [15]. (i) Adapted from [17].

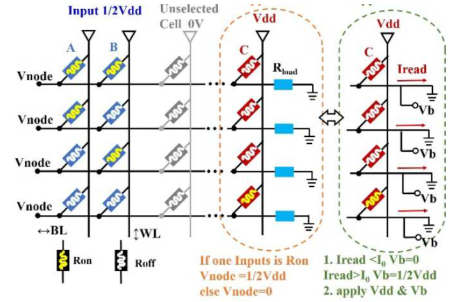
Input 1/2Vdd
Unselected Cell 0V
A
B
C
Vdd
R_{load}
Iread
Vb
Vnode
↔BL
↕WL
Ron
Roff
If one Inputs is Ron Vnode =1/2Vdd else Vnode=0
1. Iread <I_0 Vb=0 Iread>I_0 Vb=1/2Vdd
2. apply Vdd & Vb

Fig. 7. Wired-NOR logic gate implemented with binary Cu/Al2O3/poly-Si RRAM devices with low power (I_{on}<100 nA) and high on/off ratio. Adapted from [18].

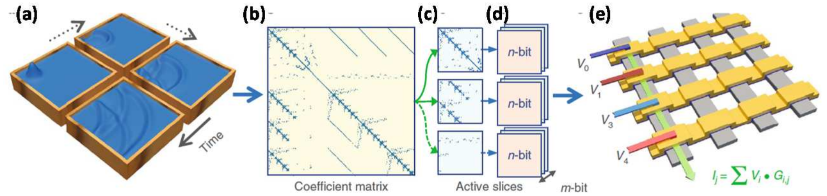

Fig. 8. High-precision PDE solver implemented in RRAM crossbar. Adapted from [19].

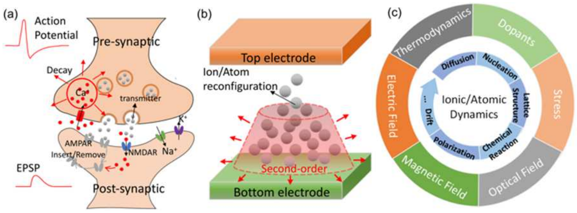

Fig. 9. (a) Synaptic plasticity can be emulated with RRAM devices through (b) internal ionic processes. (c) Coupling of internal physical processes in RRAM can lead to rich dynamic behaviors.

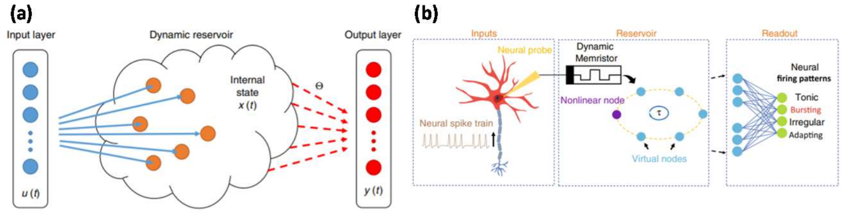

Fig. 10. Schematic of (a) an RC system consisting of an input layer, a dynamic reservoir and an output layer, and (b) an RRAM-based RC system using the multiple virtual node concept. (a) is adapted from [20]. (b) is adapted from [21].

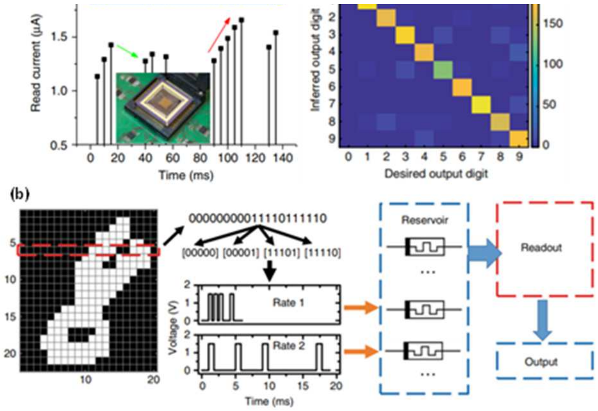

Fig. 11. (a) Response of a typical WO$_x$ RRAM device driven by temporal inputs. (b) Handwritten digit classification implemented with an RRAM-based RC system. Adapted from [21].

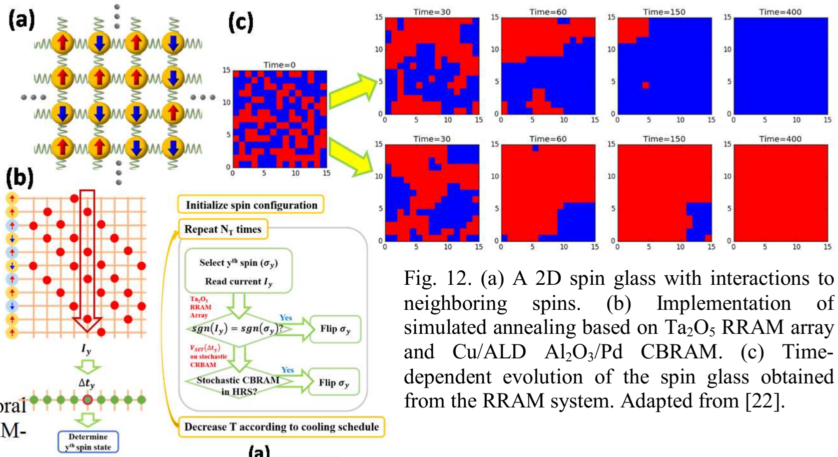
(a)
(b)
(c)
Time=0
Time=30
Time=60
Time=150
Time=400
I_y
Δt_y
Determine y^{th} spin state
Initialize spin configuration
Repeat N_T times
Select y^{th} spin (σ_y)
Read current I_y
Ta2O5 RRAM Array
$sgn(I_y) = sgn(\sigma_y)$?
Yes
Flip σ_y
$V_{SET}(\Delta t_y)$ on stochastic CBRAM
Stochastic CBRAM in HRS?
Decrease T according to cooling schedule

Fig. 12. (a) A 2D spin glass with interactions to neighboring spins. (b) Implementation of simulated annealing based on Ta_2O_5 RRAM array and Cu/ALD Al_2O_3/Pd CBRAM. (c) Time-dependent evolution of the spin glass obtained from the RRAM system. Adapted from [22].

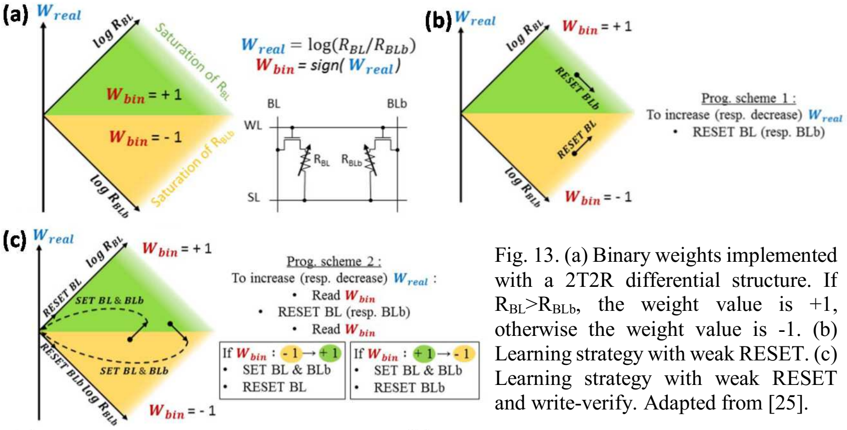

Fig. 13. (a) Binary weights implemented with a 2T2R differential structure. If $R_{BL}>R_{BLb}$, the weight value is +1, otherwise the weight value is -1. (b) Learning strategy with weak RESET. (c) Learning strategy with weak RESET and write-verify. Adapted from [25].

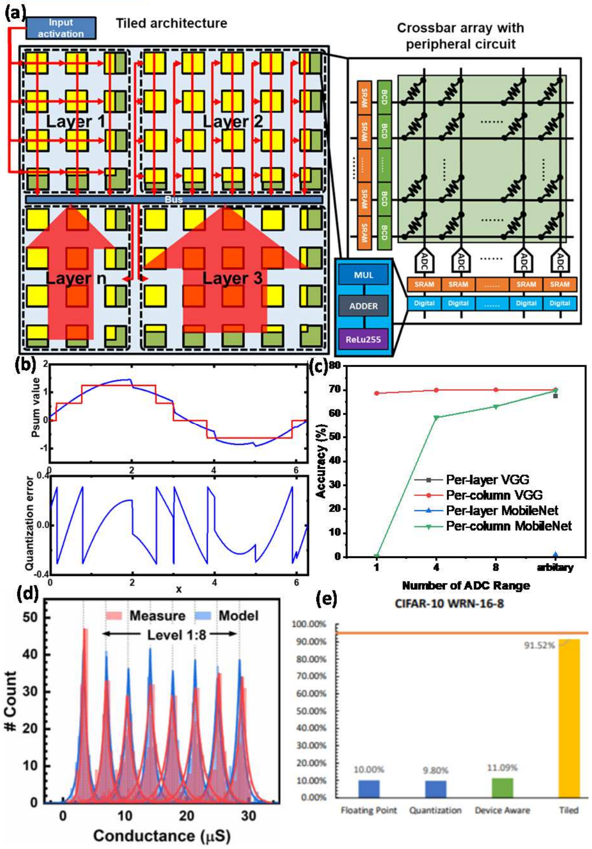

Fig. 14. (a) Tiled IMC architecture. (b) ADC quantization errors of Psum can be mitigated through (c) multi-range quantization. The influence of (d) device conductance variability and finite on-off ration can be mitigated through (e) architecture-aware training. (a) and (c) are adapted from [23]. (d) is adapted from [28]. (e) is adapted from [29].

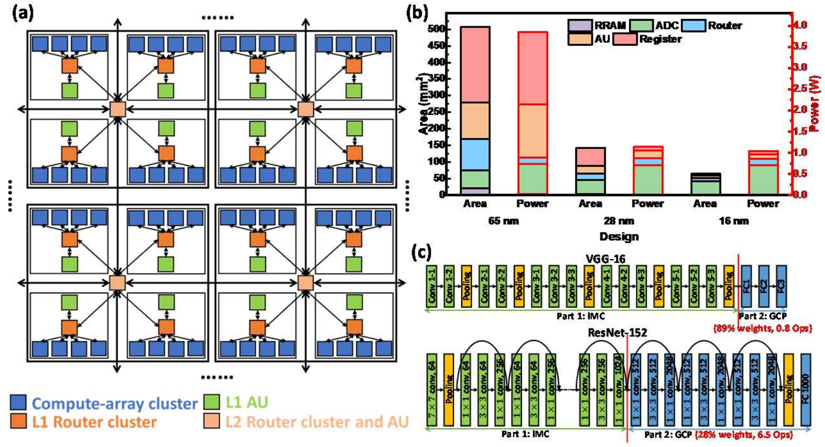

Fig. 15. (a) Schematic of the hierarchical TAICHI architecture. The HM-NoC handles data routing between two levels of clusters. (b) Power and area breakdown of the TAICHI design. (c) DNN models that exceed the storage capacity of the IMC system can be mapped onto a hybrid IMC+GCP system. (a-b) are adapted from [33].