

# ReBIT: A ReRAM-Based In-Situ Training Accelerator with Robustness Against Stochasticity

Peng Dang<sup>\*†</sup>, Wei Wang<sup>‡</sup>, Yintao He<sup>\*†</sup>, Huawei Li<sup>\*†✉</sup>

<sup>\*</sup>SKLP, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China

<sup>†</sup>University of Chinese Academy of Sciences, Beijing, China; <sup>‡</sup>Pengcheng Laboratory, Shenzhen, China

Email: dangpeng21@mailsucas.ac.cn, wangwei@pcl.ac.cn, heyintao19z@ict.ac.cn, ✉lihuawei@ict.ac.cn

**Abstract**—In-situ training architectures based on resistive random-access memory (ReRAM) have attracted significant attention due to their exceptional energy efficiency. However, the inherent stochasticity of ReRAM devices severely degrades training convergence. To address this challenge, this work proposes a ReRAM-based in-situ training accelerator (ReBIT) architecture. The ReBIT integrates ReRAM with static random-access memory (SRAM) devices, leveraging the deterministic characteristics of SRAM-based computations to suppress the inherent stochasticity of ReRAM devices. Experimental results demonstrate that the ReBIT architecture achieves convergence performance comparable to full-precision software training.

**Index Terms**—ReRAM, SRAM, Processing-in-Memory, In-Situ Training, Stochasticity

## I. INTRODUCTION

As neural network models continue to scale up, traditional computing architectures face the “memory wall” bottleneck in neural network training [1]. Processing-in-memory (PIM) architectures based on ReRAM perform matrix-vector multiplication (MVM) operations directly within memory arrays, substantially reducing data transfer overhead [2]. With their highly parallel computational capabilities and ultra-low energy consumption characteristics, these architectures provide a revolutionary solution for in-situ training [3]. However, the inherent stochasticity of ReRAM devices, including nonlinear conductance updates, cycle-to-cycle (C2C) variations, and device-to-device (D2D) variations, introduces uncertainty in weight updates that severely degrades training convergence [4, 5]. This challenge impedes the deployment of ReRAM in practical training scenarios.

In ReRAM-based in-situ training, model weights typically require low-precision quantization to match the limited conductance states of ReRAM [5]. This process employs a quantization function to transform high-precision real-valued weights  $W^{(r)}$  into low-precision quantized weights  $W^{(a)}$  [6]. The quantized weights are used to perform MVM operations during forward propagation, while the real-valued weights are utilized for matrix-transpose vector multiplication ( $M^TVM$ ) operations [3] involved in gradient computation during backward propagation. Since forward and backward propagation rely on two distinct weight representations,  $W^{(a)}$  and  $W^{(r)}$  respectively, two separate memory units are required to support the computational processes for MVM and  $M^TVM$  operations.

Based on the above analysis, this work aims to explore how to achieve highly reliable in-situ training under non-ideal hardware conditions. To this end, this work proposes

the ReBIT architecture and its accompanying cooperative computing methodology. As depicted in Fig. 1, ReBIT is a hybrid PIM architecture that combines SRAM with ReRAM. Through a strategy of forward and backward cooperative computing, ReBIT effectively suppresses random errors in analog devices, ensuring the convergence performance and accuracy of neural network training while maintaining energy efficiency. Experimental results demonstrate that, compared with conventional architectures, ReBIT achieves training accuracy comparable to software-based training under conditions with significant stochasticity, while simultaneously improving energy efficiency and accelerating the training process.

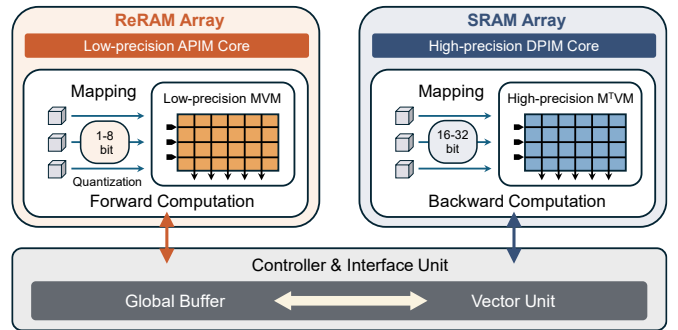


Fig. 1. Overview of ReBIT architecture.

## II. METHODOLOGY

### A. ReBIT Architecture

As illustrated in Fig. 1, the ReBIT architecture comprises multiple modules operating cooperatively, including a control unit, vector units, and global buffers. The core computing unit consists of an analog PIM (APIM) unit and a digital PIM (DPIM) unit, which are interconnected through an on-chip network to enable flexible dataflow routing. The APIM stores quantized weights and executes MVM during forward propagation, while the DPIM maintains real-valued weights and performs  $M^TVM$  required for backward propagation. Through this collaborative design, the ReBIT architecture efficiently facilitates in-situ training of neural networks.

As shown in Fig. 2(a), in the APIM architecture, a digital-to-analog converter (DAC) transforms the activations  $a$  into analog voltages that are applied to the ReRAM array, where weights  $W_{ij}$  are stored as conductance values, enabling matrix multiplication computation through physical laws. The

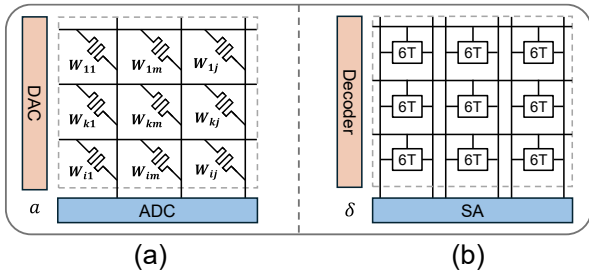


Fig. 2. (a) APIM macro; (b) DPIM macro.

computational results are then digitized through an analog-to-digital converter (ADC). As illustrated in Fig. 2(b), the DPIM architecture employs a memory array composed of 6T SRAM cells, where the error  $\delta$  is input to the SRAM array via decoders, and data sensing is performed through sense amplifiers (SAs). This structure maintains high-precision computation while providing excellent reconfigurability and stability. Both architectures embody the design philosophy of implementing neural network in-situ training at the hardware level, catering to requirements across diverse application scenarios.

### B. Stochasticity-Resilient Method for In-situ Training

To address the impact of inherent stochasticity in ReRAM, the DPIM introduced in ReBIT plays a critical role. During in-situ training, APIM and DPIM operate alternately according to the requirements of computational tasks. Specifically, in the forward propagation phase, APIM performs energy-efficient, low-precision inference computations. In contrast, DPIM handles high-precision gradient computations during the backward propagation phase. During the weight update phase, the computed weight update increments are first applied to DPIM, and subsequently synchronously mapped to APIM. This hybrid architecture integrates the respective advantages of APIM and DPIM, achieving a balance between computational energy efficiency and accuracy. In ReBIT, APIM serves as the computational engine for forward propagation, maximizing computational energy efficiency. DPIM functions as a precision anchor, responsible for high-precision gradient computation and weight updates, ensuring the entire system maintains both efficiency and accuracy throughout execution.

## III. EVALUATION

### A. Results and Discussion

This work evaluates in-situ training performance using the VGG11 on the CIFAR10 dataset. Simulations are based on TaOx/HfOx devices [7] fabricated with 32nm technology and incorporate the conductance update model proposed in [5]. To comprehensively evaluate system robustness under hardware non-idealities, both D2D and C2C variations are set to 10%. The model employs 3-bit quantized weights for forward propagation while retaining 32-bit real-valued weights for backward propagation. Fig. 3 illustrates the distribution of real-valued weights and quantized weights in the middle layer of VGG11.

Experimental results demonstrate that the baseline scheme based on a full-ReRAM architecture achieves a training accuracy of only 86.45% due to the interference of device

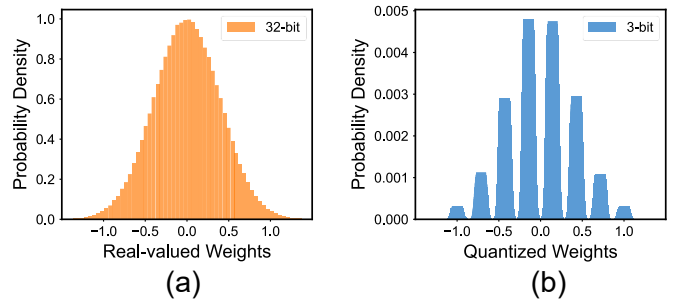


Fig. 3. Weight distribution in in-situ training. (a) Real-valued weight distribution; (b) Quantized weight distribution with 10% D2D and C2C variations.

stochasticity. In contrast, the proposed ReBIT architecture achieves a training accuracy of 91.82%, representing an absolute improvement of 5.37% over the baseline. Simultaneously, the ReBIT achieves forward computation energy efficiency of 74.29 TOPS/W and backward computation energy efficiency of 1.78 TOPS/W. These results confirm that ReBIT can achieve convergence performance comparable to full-precision software training. This performance enhancement primarily stems from the cooperative computing mechanism between DPIM and APIM: even when forward propagation is affected by hardware non-idealities, the model can compensate for hardware imperfections through precise gradient adjustments, provided that backward propagation maintains deterministic weight updates.

## IV. CONCLUSION

In this work, we propose ReBIT, a hybrid architecture that integrates ReRAM with SRAM. Through a cooperative forward and backward computing strategy, ReBIT effectively suppresses random errors in analog devices while maintaining energy efficiency. Experimental results demonstrate that this heterogeneous cooperative mechanism leverages the deterministic computation characteristics of SRAM to mask the stochasticity of ReRAM, ensuring convergence performance in in-situ training.

## ACKNOWLEDGMENT

This paper is supported in part by the Chinese Academy of Sciences under grant No. XDB0660102, and in part by the National Natural Science Foundation of China (NSFC) under grant No. 62090024.

## REFERENCES

- [1] A. Mehonic et al. "Brain-inspired computing needs a master plan". In *Nature*, vol. 604, no. 7905, pp. 255–260, 2022.
- [2] J. Liu et al. "OptiPIM: Optimizing Processing-in-Memory Acceleration Using Integer Linear Programming". In *ISCA*, pp. 867–883, 2025.
- [3] A. Ankit et al. "Panther: A programmable architecture for neural network training harnessing energy-efficient reram". In *TC*, vol. 69, no. 8, pp. 1128–1142, 2020.
- [4] W.-L. Chen et al. "A Novel and Efficient Block-Based Programming for ReRAM-Based Neuromorphic Computing". In *ICCAD*, pp. 1–9, 2023.
- [5] X. Peng et al. "DNN+NeuroSim V2.0: An End-to-End Benchmarking Framework for Compute-in-Memory Accelerators for On-Chip Training". In *TCAD*, vol. 40, no. 11, pp. 2306–2319, 2021.
- [6] S. Wu et al. "Training and inference with integers in deep neural networks". In *arXiv preprint arXiv:1802.04680*, 2018.
- [7] W. Wu et al. "A methodology to improve linearity of analog RRAM for neuromorphic computing". In *VLSI technology*, pp. 103–104, 2018.