

Data Distribution-Aware Analog/Digital Conversion Strategy for Energy-Efficient Memristive In-Situ Accelerators

Taoming Lei, Heng Zhou, Bing Wu, Wei Tong*, Dan Feng

Wuhan National Laboratory for Optoelectronics, Key Laboratory of Information Storage System, Engineering Research Center of Data Storage Systems and Technology, Ministry of Education of China, School of Computer Science and Technology, Huazhong University of Science and Technology, China
 {leitaoming, hengzhou, wubin200, tongwei, dfeng}@hust.edu.cn

*Corresponding author

Abstract—Memristive in-situ computing offers energy-efficient DNN acceleration, but faces ADC-induced energy bottlenecks. We observe that bitline outputs exhibit significant non-uniformity and cycle-to-cycle variation, rendering conventional A/D conversion schemes suboptimal. We thus propose a data distribution-aware A/D conversion strategy that predicts key bits of digital outputs and skips unnecessary steps, with a switching mechanism adapting the optimal conversion method across cycles. Implemented via a reconfigurable SAR-ADC, our approach significantly reduces the energy consumption of in-situ accelerators.

Index Terms—in-situ computing, analog/digital conversion, energy efficiency

I. INTRODUCTION

Memristive in-situ computing has demonstrated significant potential for accelerating deep neural networks (DNNs), owing to its characteristics of low power consumption and high performance. However, in-situ accelerators commonly rely on high-frequency successive approximation register analog-to-digital converters (SAR-ADCs) to convert analog signals from crossbar's bitlines into digital values. These ADCs exhibit prohibitively high energy consumption, accounting for up to 90% of the total energy budget [1], making them the primary energy bottleneck in such accelerator.

SAR-ADC conventionally employs a bit-by-bit binary search method to resolve digital values, without considering the distribution features of analog signals. However, we find that within in-situ accelerator, crossbar's bitline outputs exhibit significant non-uniformity and spatiotemporal variability, as shown in Fig. 1. Therefore, we present a novel architecture based on a reconfigurable SAR-ADC design to reduce A/D conversion energy. This design implements the data distribution-aware A/D conversion strategy, which predicts higher-order bits of digital outputs based on crossbar's bitline output distribution features, thereby skipping unnecessary conversion steps. We customize distinct A/D conversion methods for different distribution types

This research was supported in part by the Young Scientists Fund of the National Natural Science Foundation of China under Grant 62302179, the National Natural Science Foundation of China under Grant 62172178, the Young Scientists Fund of the National Natural Science Foundation of China under Grant 62302182, and Cross-Research Support Program of Huazhong University of Science and Technology under Grant 2023CJYJ042.

and introduce a switching mechanism that adaptively selects the suitable conversion method at runtime based on the varying bitline output distributions. Experiments show that our approach significantly reduces the energy consumption of in-situ accelerators.

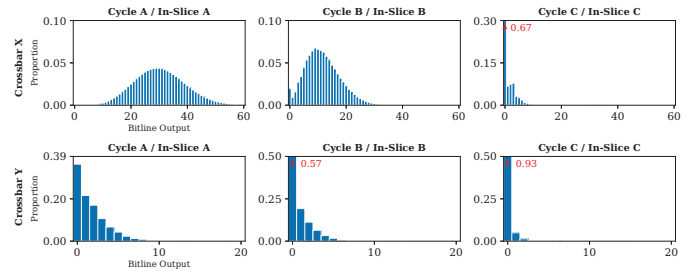


Fig. 1. Bitline Output Distributions of Two Crossbars. Each row of subplots corresponds to a different crossbar. Each column of subplots displays the output distribution under a different activation bit-slice input across cycles. Red triangles represent excessively large outliers.

II. DATA DISTRIBUTION-AWARE ANALOG/DIGITAL CONVERSION STRATEGY

A. A/D Conversion Methods for Different Distribution

The bitline outputs of crossbar in memristive in-situ accelerators mainly follow normal or monotonically biased distributions, as shown in Fig. 1. Accordingly, we design dedicated A/D conversion methods tailored to each distribution type, beginning with the Conversion Method for Biased distributions (CMB). In such distributions, most higher-order bits are zero, enabling prediction-based step skipping. CMB uses two parameters, N_{start} and N_{step} , to control the conversion starting point and handle prediction failures, as depicted in Fig 2a.

For an R -bit ADC, CMB predicts the first N_{start} high-order bits as zero and generates the initial reference code $S_{ref} = 2^{R-1-N_{start}}$. S_{ref} is converted to V_{ref} for comparison with V_{in} . If $V_{in} < V_{ref}$, prediction succeeds. Otherwise, CMB rolls back N_{start} by N_{step} bits and retries prediction until success or no bits remains.

For normally distributed outputs, we propose the Conversion Method for Normal distributions (CMN), which introduces an offset parameter N_{off} to start near the distribution peak, as

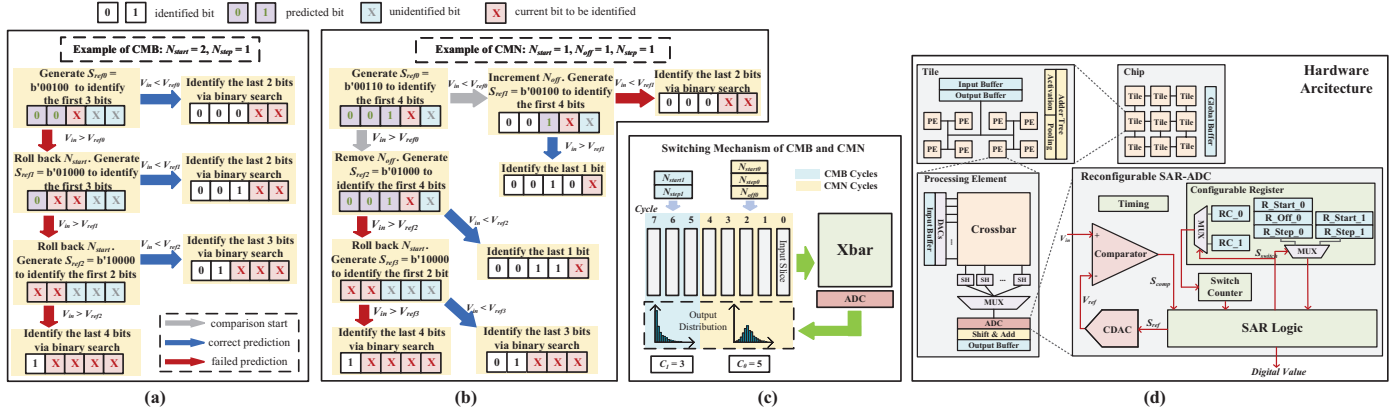


Fig. 2. (a) Example of A/D Conversion Method for Biased Distribution (CMB). (b) Example of A/D Conversion Method for Normal Distribution (CMN). (c) Switching mechanism of different conversion methods. (d) Proposed architecture based on the reconfigurable SAR-ADC design.

depicted in Fig. 2b. The initial reference is set as $S_{ref} = 2^{R-1-N_{start}} - 2^{N_{off}}$. If $V_{in} < V_{ref}$, CMN increments N_{off} by N_{step} and iterates until V_{in} lies between two successive reference values, thereby determining the higher-order bits. If $V_{in} > V_{ref}$, CMN falls back to CMB by removing N_{off} and proceeding accordingly.

B. Switching Mechanism of A/D Conversion Methods

We observe that the bitline output distributions vary with different input slice per cycle, as shown in Fig. 1. Rather than configuring parameters per slice, we group outputs by distribution type and allow each group to use one configuration to cut register overhead, i.e., outputs of normal distributions share one parameter set and biased distributions another.

Two cycle counters, C_0 and C_1 , control switching between CMN and CMB. As depicted in Fig. 2c, with 8 input slices, cycles 0 to 4 that follow a normal distribution use $C_0 = 5$ and share parameters $N_{start0}, N_{step0}, N_{off0}$. Cycles 5 to 7, which have a biased distribution, use $C_1 = 3$ and share parameters N_{start1}, N_{step1} . The system switches back to CMN after cycle 7 and repeats the procedure. This enables distribution aware dynamic switching with minimal parameter storage.

We use an offline search to find parameters that minimize conversion steps for CMB and CMN. The algorithm iterates over all possible combinations. We sample 50 training images, run the model, and record each crossbar's output distribution. For each crossbar, the optimal set is the combination that yields the fewest total steps.

C. Architecture Design

We propose a novel in-situ computing architecture based on a reconfigurable SAR-ADC design as shown in Fig. 2d. This design supports the data distribution-aware A/D conversion strategy through a set of configurable registers and additional control logic. Specifically, registers R_Start_0, R_Step_0 and R_Off_0 store the parameters $N_{start0}, N_{step0}, N_{off0}$ for CMN, while registers R_Start_1 and R_Step_1 store the parameters N_{start1}, N_{step1} for CMB. Registers RC_0 and RC_1 stores the values of variables C_0 and C_1 , which determine the processing periods of CMB and CMN as described in Section II-B. A

switch counter controls the dynamic switching between the different conversion methods.

III. EVALUATION

We modify MNSIM-2.0 [2] to evaluate the accelerator architectures. The parameters of ADC are obtained from [3]. Digital component of peripheral circuit are synthesized using Design Compiler. For comparison, we select ISAAC [4] and TRQ [5] as our baseline.

Fig. 3 shows the overall energy breakdown per processed image for the four models mentioned above. Compared to ISAAC, our approach reduces ADC energy consumption by 76.46%, 73.07%, 75.56% and 75.49% for AlexNet, VGG-11, VGG-16, and ResNet-18, respectively. When compared to TRQ, it still achieves additional reductions of 41.14%, 40.83%, 41.81% and 41.60% for these same workloads. Notably, the added registers and control logic increase ADC area by 7.59% and overall tile area by only 3.97%. Given the substantial energy savings, this overhead is acceptable.

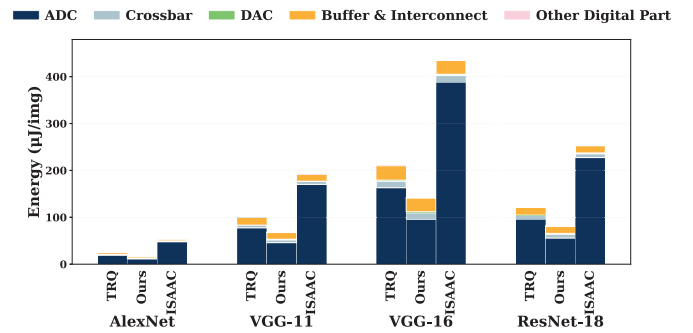


Fig. 3. Energy breakdown of memristive in-situ accelerator.

IV. CONCLUSION

In this paper, we propose a data distribution-aware A/D conversion strategy to reduce ADC energy in memristive in-situ accelerators. Our approach predicts higher-order bit values from bitline distributions to skip unnecessary steps, and incorporates a dynamic switching mechanism to adapt optimal conversion methods across cycles. Experiments show that our approach significantly reduces the energy consumption of in-situ accelerators.

REFERENCES

- [1] X. Li, Z. Yuan, et al., "Tailor: Removing redundant operations in memristive analog neural network accelerators," in *DAC*, 2022, pp. 1009–1014.
- [2] Z. Zhu, H. Sun, et al., "Mnsim 2.0: A behavior-level modeling tool for processing-in-memory architectures," *TCAD*, pp. 4112–4125, 2023.
- [3] L. Kull, T. Toifl, et al., "A 3.1 mw 8b 1.2 gs/s single-channel asynchronous sar adc with alternate comparators for enhanced speed in 32 nm digital soi cmos," *JSSC*, pp. 3049–3058, 2013.
- [4] A. Shafiee, A. Nag, et al., "Isaac: A convolutional neural network accelerator with in-situ analog arithmetic in crossbars," in *ISCA*, 2016, pp. 14–26.
- [5] C. Zhang, Z. Yuan, et al., "Algorithm-hardware co-design for energy-efficient a/d conversion in reram-based accelerators," in *DATE*, 2024, pp. 1–6.