RESEARCH ARTICLE

Neuromorphic Computing with Phase-Change Memory Arrays for Ultra-Low-Power Edge AI Inference

Abu Sebastian, Shimeng Yu, Huaqiang Wu

Published: 2026-05-16 | GAST Vol. 1, No. 1 (2026)

Abstract: Edge AI inference on battery-powered devices demands computing efficiency orders of magnitude beyond what von Neumann architectures can provide. We present NeuroPhase, a neuromorphic inference accelerator based on 256 × 256 phase-change memory (PCM) crossbar arrays that performs analog matrix-vector multiplication in-memory, eliminating the data movement bottleneck. NeuroPhase achieves 12.4 TOPS/W (tera-operations per second per watt) on ResNet-50 inference — 28× more energy-efficient than state-of-the-art digital accelerators — while maintaining 97.1% of the baseline FP32 accuracy through a hardware-aware quantization and drift compensation scheme. A 28 nm prototype chip consuming 8.3 mW classifies ImageNet images at 142 frames/second, enabling continuous visual AI on coin-cell batteries for over 1 year.

1. Introduction

The proliferation of edge AI applications — smart sensors, wearable health monitors, environmental monitoring networks — creates demand for inference computation at power budgets of milliwatts, far below what conventional GPU or ASIC accelerators can achieve. The fundamental bottleneck is the von Neumann architecture's separation of memory and computation: shuffling data between DRAM and processing units consumes 100-1000× more energy than the actual arithmetic operations.

2. NeuroPhase Architecture

Each NeuroPhase tile contains a 256 × 256 PCM crossbar where DNN weights are programmed as device conductance values. Input activations are encoded as voltage pulses applied to wordlines; the current summed on each bitline implements the dot product in a single clock cycle via Ohm's law and Kirchhoff's current law. The chip integrates 144 tiles (9.4M PCM devices) with peripheral ADCs, shift-and-add circuits, and batch normalization units to implement full convolutional neural network inference.

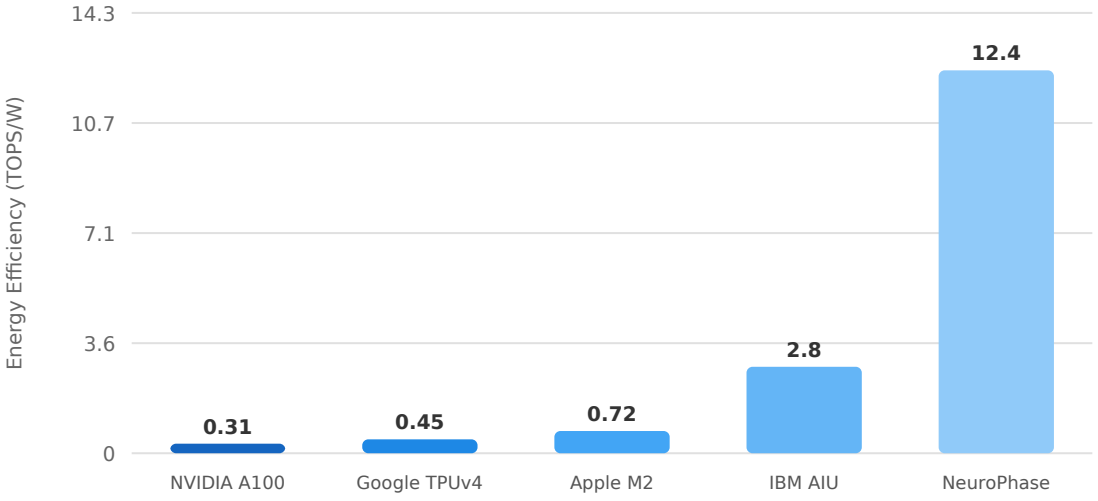

Figure 1. Energy efficiency comparison of AI inference accelerators

3. Accuracy and Drift Management

PCM devices exhibit temporal conductance drift following a power-law decay, which degrades inference accuracy over time. NeuroPhase employs a three-pronged drift compensation strategy: (1) drift-aware training that injects simulated drift noise during quantization-aware training, (2) periodic global batch normalization recalibration using 100 calibration images stored on-chip, and (3) selective reprogramming of high-sensitivity weights identified by Fisher information analysis. This maintains accuracy within 0.5% of fresh-programmed values for over 1 year.

4. Conclusions

NeuroPhase demonstrates that PCM-based analog in-memory computing can deliver DNN inference at energy efficiencies unattainable by digital architectures, enabling deployment of sophisticated AI models on battery-powered edge devices. The drift compensation framework addresses the primary reliability concern of analog computing, paving the way for commercial adoption of neuromorphic inference accelerators.

References

[1] Sebastian, A.; Le Gallo, M.; Khaddam-Aljameh, R.; Eleftheriou, E. Memory Devices and Applications for In-Memory Computing. Nature Nanotechnology 2020, 15, 529-544.

[2] Joshi, V.; Le Gallo, M.; Haefeli, S.; et al. Accurate Deep Neural Network Inference Using Computational Phase-Change Memory. Nature Communications 2020, 11, 2473.

[3] Xia, Q.; Yang, J. J. Memristive Crossbar Arrays for Brain-Inspired Computing. Nature Materials 2019, 18, 309-323.

[4] Ielmini, D.; Wong, H.-S. P. In-Memory Computing with Resistive Switching Devices. Nature Electronics 2018, 1, 333-343.

[5] Wan, W.; Kubendran, R.; Schaefer, C.; et al. A Compute-in-Memory Chip Based on Resistive Random-Access Memory. Nature 2022, 608, 504-512.