A Via-Programmable Neuron Array for Wearable AI-IoT Applications

UTokyo, ◦ Jaewon Shin, Rei Sumikawa, Dongzhu Li, Mototsugu Hamada, Atsutake Kosuge

E-mail: jwshin@g.ecc.u-tokyo.ac.jp

1. Introduction

Wearable AI-IoT devices have become essential in areas such as health monitoring, environmental sensing, and activity tracking. These applications require processors that operate efficiently under strict constraints of power, area, and cost. Conventional FPGA-based solutions, though flexible, are energy-inefficient and unsuitable for continuous operation. On the other hand, task-specific ASICs, while optimized for performance, incur high manufacturing costs due to their application-specific design, requiring another full set of masks. To address these limitations, we propose a via-programmable neuron array (VPNA) architecture (Fig. 1). This design offers a low-cost and energy-efficient solution for diverse AI-IoT tasks.

2. Proposed Methods

The VPNA consists of a 64 × 64 tile-based array optimized for one-dimensional neural networks, commonly used in time-series data analysis. To support diverse applications, each VPNA tile is configurable into various types of 1D neural network layers, including 1D convolution, max pooling, and median filter. By concatenating multiple tiles, the architecture enables the implementation of complex DNN models. Each tile processes 16-bit data and supports ternary weights (+1, 0, -1), which are implemented as programmable vias. These vias allow specific DNN models to be hardwired by customizing their locations based on DNN training results. As a result, customized DNN processors can be fabricated with the same mask set, except for a single via mask for programming.

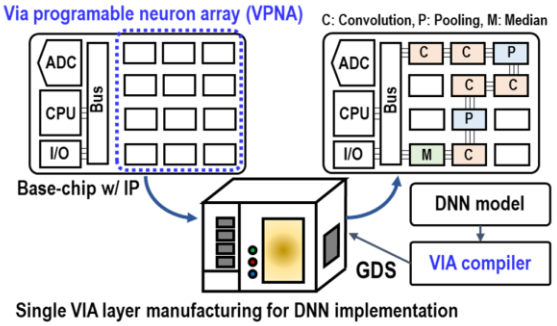

Fig. 1 Via-programming DNN processor fabrication

The technical challenge for VPNA is the large number of signal wires. The bit- and neuron-serial circuit (BNSC) reduces wiring complexity by a factor of 1,024, leveraging bit-serial communication and a multi-drop bus configuration. Additionally, the function-selective non-linear neural network (FS-NNN) simplifies the implementation of non-linear activation functions by limiting them to four specific types: ReLU, inverse ReLU, linear, and inverse functions. This simplification reduces the hardware resources needed. Together, these techniques enhance the scalability and practicality of the VPNA for diverse use cases.

3. Results

We fabricated a test chip with VPNA tiles using a 40nm CMOS process. To evaluate its performance, the VPNA was configured as a 1D convolution layer with 50% weight sparsity. The results demonstrated that the VPNA achieves comparable or superior energy efficiency across various applications while utilizing the same base chip. When compared with an FPGA-based KWS recognition processor [1], the VPNA achieved 28.3 times higher energy efficiency. Furthermore, despite being implemented in an older 40nm process, the VPNA reduced the area by a factor of 2.3 compared to the 14nm analog reconfigurable CiM [2], which has a similar tile structure and reconfigurable wiring. These improvements highlight the effectiveness of the VPNA's design, particularly the contributions of BNSC and FS-NNN, in optimizing both energy efficiency and area utilization.

References

[1] H. Borras et al., “Open-source FPGA-ML codesign for the MLPerf Tiny Benchmark,” MLSys., pp. 1-15, Aug 2022

[2] S. Ambrogio et al., “An analog-AI chip for energy-efficient speech recognition and transcription,” Nature 620, 768-775 (2023)