ZE-WEI (Johnny) LIOU

Email: zwljohnny@gmail.com *LinkedIn* *Ze-Wei Liou's Webpage*

INTERESTS

AI Accelerator Design, Compute-in-Memory (CIM) Systems, Multicore AI Accelerators, Dynamic Inference, N:M Sparse, Transformer Quantization, VLSI System Design

EDUCATION

Sep 2020 - Jun 2024 National Taiwan University Department of Electrical Engineering

- **Overall GPA: 4.12/4.3 (Last 60 GPA: 4.25/4.3)**
- **Ranked 1st in 3 semesters (1st among 262 in senior year)**

PUBLICATION

[1] **Liou, Z.-W.**, Hong, D.-Y., "Optimizing Compute Core Assignment for Dynamic Batch Inference in AI Inference Accelerator," ACM Symposium on Applied Computing (SAC), 2025. (*Under Review)*

[2] Burr, G. W., Tsai, H., Simon, W., Boybat, I., Ambrogio, S., Ho, C.-E., **Liou, Z.-W.**, et al., "Design of Analog-AI Hardware Accelerators for Transformer-based Language Models (Invited)," IEDM, 2023.

RESEARCH EXPERIENCE

Research Assistant | Academia Sinica **Feb 2024 - Present**

Advised by Dr. Jan-Jan Wu and Dr. Ding-Yong Hong

Efficient Dynamic Batch Inference on Multicore AI Accelerators [1]

- Proposed a dynamic programming algorithm for optimizing compute core assignments on multicore AI accelerators to maximize throughput of dynamic batch inference workloads.
- Achieved 3.05x higher inference throughput on benchmark datasets compared to the EdgeTPU inference strategy.

Flexible Hardware Accelerator for N:M Sparse Models

- Designed a flexible systolic array-based hardware accelerator capable of supporting arbitrary N:M sparsity ratios for both activations and weights on the Xilinx ZCU104 FPGA.
- Developed an intra-matrix block-wise N:M sparsity scheme supported by flexible hardware, achieving a 2% accuracy improvement on BERT models across several GLUE benchmarks compared to layer-wise N:M sparsity.

Research Intern | IBM Research Almaden Lab, San Jose, CA **Jun 2023 - Sep 2023**

Advised by Dr. Geoffrey Burr and Dr. HsinYu (Sidney) Tsai

Efficient Compute-in-Memory (CIM) Architecture for Encoder-Based Transformer Models

- Analyzed system-level design trade-offs, including latency and area, between volatile and non-volatile memory-based CIM systems on BERT models. [2]
- Designed highly pipelined volatile memory-based architectures. Identified design choices, such as the number of Analog Fabrics, that optimized area efficiency (TOPS/sec/sq.mm) across various input scenarios.
- Implemented functional verification in a C++-based CIM simulator to automatically verify the correctness of data transfers and computational results.

Undergraduate Researcher | NTU Access IC Lab **Jul 2022 - Jul 2023**

Advised by Prof. An-Yeu (Andy) Wu

CIM-aware Algorithm for Vision Transformer Quantization

- Addressed accuracy drop in quantized Vision Transformers caused by severe inter-channel variation among LayerNorm inputs by applying a CIM-aware group quantization approach to efficiently handle outlier data.
- Designed RRAM-based CIM architectures on a C++-based CIM simulator that support the group quantization method. Reduced the inference energy of the LayerNorm layer by 25% while preserving accuracy.

AWARDS

National Integrated Circuit Design Contest - Second Prize **May 2023**

National Taiwan University Dean's List Award **Spring 2023**

National Meichu Hackathon - First Prize **Oct 2021**

VLSI CIRCUIT DESIGN EXPERIENCE

2D Bilinear Resize Engine **May 2023**

Second Prize Award in National Integrated Circuit Design Contest

- Implemented by designing a floating-point bilinear interpolation computation element.
- Conducted RTL coding and logic synthesis, optimizing performance through pipelined implementation.

5G MIMO Demodulation **May 2023 - Jun 2023**

Project Designed by MediaTek and National Taiwan University

- Implemented maximum likelihood demodulation in a MIMO receiver through the complete VLSI system design flow, including RTL design, logic synthesis, and placement.
- Achieved second-best area × power × time performance in post-layout design among 44 graduate groups.

Real Time FPGA-based Acoustic Imaging **Nov 2022 - Dec 2022**

- Implemented a hardware system with a beamforming algorithm on an Altera Cyclone IV FPGA to visualize the location and intensity of sound waves.

Courses

- Achieved A+ grades in several VLSI courses, including Computer-Aided VLSI System Design, Integrated Circuit Design, Digital Signal Processing in VLSI, Digital Circuit Lab, and Computer Architecture.

HACKATHON EXPERIENCE

E-Air: Intelligent air conditioner system **Oct 2021**

First Prize Award in National Meichu Hackathon

- Obtained human pose data using a model from MediaPipe and then adjusted the speed and swing of the air conditioner model we created.
- Developed strong problem-solving skills and effectively implemented solutions in practice.

VOLUNTEER WORK AND AFFILIATIONS

Public Relations Officer | School‘s Country Youth Service Club **Sep 2021 - Present**

- Shared scientific knowledge, including topics such as air pressure and magnetic force, with rural junior high school students through games during each semester's camp.
- Efficiently communicated with service-oriented foundations and rural schools.