Charlie (Chia-Yi) Chin

Phone: (+886) 966-332-562 Email: cychin424@gmail.com LinkedIn: Chia-Yi Chin

Education

National Taiwan University (NTU) **Taipei, Taiwan**
B.S. in Mechanical Engineering (transferred tracks) Sep 2021 – Jun 2022
B.S. in Electrical Engineering Sep 2022 – Jan 2026 (expected)
GPA: 4.12 / 4.3, Last 60: 4.22 / 4.3

Research Experience

Elsa Lab, NTU; Advisor: Chun-Yi Lee **Taipei, Taiwan**
Research Topic: **LLM Distributed Inferences Acceleration** Feb 2025 – Present
- Developed workload distribution algorithm on an Intel CPU cluster using MPI and OpenMP
- Leveraged AVX-512 and AMX ISA to accelerate vector and matrix operations in the Transformer architecture
- Analyzed techniques such as model quantization, KV cache compression, and kernel fusion to reduce latency and increase throughput, achieving a 25x improvement compared to the baseline version on GPU cluster

Digital Circuits and Systems Lab, NTU; Advisor: Chia-Hsiang Yang **Taipei, Taiwan**
Research Topic: **AI Accelerator Design with High-Bandwidth Memory in FPGA** Jan 2024 – Jun 2025
- Developed efficient dataflow for the Transformer model to maximize memory bandwidth utilization
- Integrated a customized IP with an ARM CPU into a SoC system, enabling flexible operation for different AI models
- Modeled AI SoC accelerator performance using HBM behavioral simulation to analyze latency and throughput

Digital Image and Signal Processing Lab, NTU; Advisor: Jian-Jiun Ding **Taipei, Taiwan**
Research Topic: **Time-Frequency Analysis of Music Signal with Color Noise** Mar 2024 – Dec 2024
- Utilized Gabor transform to analyze time-frequency spectrum distribution of a noisy signal
- Developed dynamic masking mechanism and Wiener filter to eliminate color noise effects
- Achieved high SNR with an improvement of up to 70% compared to the original signal

Academic Experience

Computer-aided VLSI System Design: Elliptic Curve Ed25519 Cryptographic Processor Nov 2024 – Dec 2024
- Implemented efficient algorithms, including extended coordinates, Karatsuba multiplication, and binary modular inversion, to accelerate point scalar multiplication
- Completed the entire design flow, including simulation, synthesis, place & route, to generate final GDSII files
- Final specs: Layout chip area of 2.56 mm², operating at a frequency of 125 MHz using TSMC 0.18 μm technology

Digital System Design: Pipelined RISC-V CPU May 2024 – Jun 2024
- Designed 5-stage pipelined multi-cycle CPU with dynamic branch prediction and hazard detection
- Implemented decompressor module to support RISC-V compressed instruction extensions

Parallel Programming: Parallelization of Multi-Key RSA Cryptography Nov 2024 – Dec 2024
- Utilized the CGBN (CUDA Accelerated Multiple Precision Arithmetic using Cooperative Groups) library to distribute threads across different workloads efficiently
- Executed modular exponentiation in parallel on the GPU Streaming Multiprocessor
- Achieved over 90% GPU speedup compared to the CPU for key lengths over 2048 bits

Digital Communication IC Design: MIMO Sphere Decoding Processor, Artix-7 FPGA Nov 2024 – Dec 2024
- Designed a pipelined hardware architecture for breadth-first sphere decoding
- Leveraged the K-best algorithm to improve throughput and reduce decoding latency

Digital Circuit Lab: Digital Audio Recorder and Player, Altera DE2-115 FPGA Oct 2023 – Dec 2023
- Utilized I2C protocol to initialize and communicate with WM8731 audio codec
- Implemented algorithms like downsampling, upsampling, and interpolation for audio data processing

Competition & Conference Experience

2025 Student Cluster Competition (SCC25), St. Louis, MO, USA Jun 2025 – Nov 2025 (expected)
- Focusing on optimizing the HPL-Mixed Precision (HPL-MxP) benchmark through system-level tuning
- Using MPI to parallelize SST (Structural Simulation Toolkit) simulations, exploring innovations in highly concurrent systems across ISA, microarchitecture, memory, and communication models

2025 APAC HPC-AI Competition (HPC-AI) Jun 2025 – Oct 2025 (expected)
- Exploring system-level optimization of DeepSeek-R1 using the SGLang framework on 2 nodes with 16 H100 GPUs
- Analyzing distributed inference system performance using NCCL communication, evaluated with NCCL benchmark

- Optimizing throughput through techniques such as Tensor Parallelism (TP), Expert Parallelism (EP), Multi-token Prediction (MTP), and the DeepGEMM kernel library

2025 ASC Supercomputer Challenge (ASC25), Qinghai, China Jan 2025 – May 2025

- Achieved First Prize and Group Competition Award
- Optimized HPL benchmark performance through high GPU utilization and system-level tuning
- Reducing latency of AlphaFold3 model by parallelizing the denoising process in diffusion models across five nodes
- Improved Transformer inference by applying data parallelism and StreamingLLM to reduce latency and increase throughput

Awards & Honors

- Selected for the Dean's List during Freshman Year and Senior Fall Semester
- Achieved 1st rank in NTUEE during the Fall 2024 semester

Knowledge & Skills

Software Programming Languages:

- Python (PyTorch, Numpy, Pandas, OpenCV), MATLAB
- C / C++ (Pthread, OpenMP, OpenMPI), GPU Programming (CUDA, HIP, Triton)
- Open-source LLM serving frameworks: TensorRT-LLM, vLLM, SGLang, and llama.cpp

Hardware Description Language & CAD Tools:

- Verilog / SystemVerilog, SystemC, Vivado, Catapult-HLS, Stratus-HLS
- VCS, Verdi, Design Compiler, PrimeTime, Innovus, IC Compiler I & II

VLSI Design Skills

- SoC Design: AMBA Bus Protocol (APB, AHB, AXI), Clock-Domain-Crossing (CDC)
- Low Power Design: Clock-Gating, Unified Power Format (UPF), IR Drop Power Analysis
- Verification & Testing: Formal Verification, Design for Testability (Scan Chain), Memory Behavior Modeling
- Advanced Process Experience: Designed digital circuits using ADFP-TSMC 16nm virtual process

Languages: Mandarin: Native proficiency | English: Intermediate proficiency

Extracurricular

NTUEE Camp Sep 2023 – Sep 2024
Director

- Led 5-member team to develop RPG game by managing task progress and coordinating the overall timeline
- Formulated schedule that aligned individual responsibilities and held regular check-ins to ensure steady progress
- Conducted regular weekly online discussions to monitor project progress and proactively resolve challenges

NTU Badminton Club Sep 2024 – Dec 2024
Secretary

- Maintained comprehensive member profiles and recorded attendance for weekly club meetings to track participation and engagement
- Managed official documents and files to ensure organizational compliance and facilitated smooth communication across departments