Yang, Chien-Yi

☎ +1(858)-241-7536 ✉ chy036@ucsd.edu github.com/doctry

Objective

4th-year CS Ph.D. candidate at UCSD looking for research-oriented intern positions with experience in compiler optimization for processing-in-memory (PIM) architectures, and large-language-model (LLM) training.

Education

University of California San Diego **September 2022 – Present**
Ph.D. in Computer Science and Engineering

National Taiwan University **September 2017 – June 2021**
Bachelor of Electrical Engineering

Work Experience

MediaTek **December 2021 – June 2022**
R&D Scientist

- Deployed a distributed training system of an LLM (T3-based model) on a cluster of 100+ GPUs with **PyTorch Distributed** using data parallel and model parallel techniques. Training includes pretraining and fine-tuning.
- Sped up the training system by 30x by careful profiling and reducing the critical path using **PyTorch Profiler**.
- Designed a **FAISS** information retrieval system by augmenting transformers in **PyTorch** to reduce error rate by 40%.

Research Experience

University of California San Diego **September 2022 – Present**
PhD student

- Designed a multi-objective software-hardware co-optimization algorithm with constraints using **BoTorch** for processing-in-memory (PIM) that improves the constraint satisfaction rate by 3.3x, increasing the accuracy and PPA by 4.28% accuracy improvement, 35.38% power reduction, 49x speedup, and 10% area reduction.
- Built the PIM design simulator using **Timeloop** for ASIC, **CiMLoop** for PIM, and **PyTorch** for noise-aware training.
- A PIM-specific compiler for DNN workloads. Reduced the number of cycles by 20x on a realistic PIM model (Samsung HBM-PIM) using a mixed-integer linear programming (MILP)-based optimization technique with **Gurobi**.
- Fast backend evaluation design space exploration (DSE) for PIM design using LLM RTL generation and hierarchical P&R design flow with **Design Compiler** for synthesis and **Innovus** for P&R.
- A hierarchical NN model to solve Steiner tree problem. Built a customized distributed training system using **PyTorch Distributed** that achieves > 98% GPU utilization on all GPU instances.

Publications

- **C.-Y. Yang**, J. Liu, M. Zhou, L. Josipovic, T. Rosing. FastPIM: LLM-Enabled Fast Backend Design for Processing-in-Memory. **ASP-DAC'26** (under review).
- J. Liu, M. Zhou, Y. Pan, **C.-Y. Yang**, L. Josipovic, T. Rosing. OptiPIM: Optimizing Processing In-Memory Acceleration Using Integer Linear Programming. **ISCA'25.**
- **C.-Y. Yang**, M. Zhou, F. Ponzina, S. S. Prakash, R. Ayoub, P. Mercati, M. Subedar, T. Rosing. Multi-objective software-hardware co-optimization for HD-PIM via noise-aware Bayesian optimization. **ICCAD'24**.
- Kahng, A. B., R. R. Nerem, Y. Wang, and **C.-Y. Yang**. NN-Steiner: A mixed neural-algorithmic approach for the rectilinear Steiner minimum tree problem. **AAAI'24**. (**Alphabetical Order**)
- Cheng, C.-Y., **C.-Y. Yang**, Y.-H. Kuo, R.-C. Wang, H.-C. Cheng, and C.-Y. Huang. Robust qubit mapping algorithm via double-source optimal routing on large quantum circuits. **TQC'24**.
- Chang, F.C., Tseng, Y.W., Yu, Y.W., Lee, S.R., Cioba, A., Tseng, I.L., Shiu, D.S., Hsu, J.W., Wang, C.Y., **Yang, C.Y.** and Wang, R.C. Flexible chip placement via reinforcement learning: late breaking results. **DAC'22**.

Projects

Qsyn
An open-sourced quantum circuit compilation framework ***100+ stars***

- A qubit mapping framework that scales up to 20,000 qubits. Integrated **PyTorch** in **C++** framework.

Skills

Programming Languages: Python, C++, Rust, Tcl, JavaScript, Java, Matlab
Technologies: Git, Shell Script, Docker, Kubernetes, PyTorch, Design Compiler, Innovus
Languages: Chinese, English, Japanese, Spanish