ZHONGKAI YU

Phone: 6199538219 | Email: zhy055@ucsd.edu | Web: https://zhongkaiyu.github.io/

Education

2024 – Now **Ph.D in Computer Science and Engineering**, University of California, San Diego (UCSD), USA
Advisor: Prof. Yufei Ding

2021 - 2024 **M.E. in Computer Technology**, Chinese Academy of Sciences (ICT, CAS), China
Advisor: Prof. Yunji Chen

2017 – 2021 **B.S. in Physics** (Zhiyuan Honors Program), Shanghai Jiao Tong University (SJTU), China

2019 – 2019 **Exchange Student** in Physics / EECS, Massachusetts Institute of Technology (MIT), USA

Research Interest

My research centers on the co-evolution of LLM and hardware systems: I explore novel architectures and systems to accelerate LLM workloads, while harnessing LLMs to advance chip design. My work lies at the intersection of Computer Architecture, ML Systems, and AI4Chip.

Publications

ISCA' 26	**Zhongkai Yu**, Yue Guan, Zihao Yu, Chenyang Zhou, Zhengding Hu, Shuyi Pei, Yangwook Kang, Yufei Ding, Po-An Tsai. "Patterns behind Chaos: Forecasting Data Movement for Efficient Large-Scale MoE LLM Inference" 🏆 **Best Paper Award** (2/845)
MICRO' 24	**Zhongkai Yu***, Shengwen Liang*, Tianyun Ma, Yunke Cai, Ziyuan Nan, Di Huang, Xinkai Song, Yifan Hao, Jie Zhang, Tian Zhi, Yongwei Zhao, Zidong Du, Xing Hu, Qi Guo, Tianshi Chen. "Cambricon-LLM: A Chiplet-Based Hybrid Architecture for On-Device Inference of 70B LLM"
MICRO' 26	**Zhongkai Yu**, Haotian Ye, Chenyang Zhou, Ohm Rishabh Venkatachalam, Zaifeng Pan, Zhengding Hu, Junsung Kim, Won Woo Ro, Po-An Tsai, Shuyi Pei, Yangwook Kang, Yufei Ding. "AMMA: A Multi-Chiplet Memory-Centric Architecture for Low-Latency 1M Context Attention Serving"
[Submitted]	**Zhongkai Yu**, Ohm Rishabh Venkatachalam, Zheng Wang, Yikai Li, Yichen Lin, Zihao Yu, Yuke Wang, Liu Liu, Xulong Tang, Shuyi Pei, Yangwook Kang, Yufei Ding. "CLoRA: A CXL-Based System for Cost-Efficient Multi-LoRA Serving"
ArXiv [Submitted]	**Zhongkai Yu***, Chenyang Zhou*, Yichen Lin, Hejia Zhang, Haotian Ye, Junxia Cui, Zaifeng Pan, Jishen Zhao, Yufei Ding. "ChipBench: A Next-Step Benchmark for Evaluating LLM Performance in AI-Aided Chip Design"
ArXiv [Submitted]	**Zhongkai Yu***, Yichen Lin*, Chenyang Zhou, Yuwei Zhang, Kun Zhou, Junxia Cui, Haotian Ye, Zhengding Hu, Ruiyi Wang, Yujie Zhao, Hejia Zhang, Jingbo Shang, Jishen Zhao, Yufei Ding. "ChipMATE: Multi-Agent Training via Reinforcement Learning for Enhanced RTL Generation"

TACO [Submitted]	Zheng Wang*, **Zhongkai Yu***, Kaijian Wang, Yichen Lin, Yikai Li, Liu Liu, Xulong Tang, Yuke Wang, Yangwook Kang, Yufei Ding. “ReDM: Efficient Training System for Large-Scale Recommendation Models on Disaggregated Memory”
OSDI’ 25	Yue Guan, Yuanwei Fang, Keren Zhou, Corbin Rebeck, Manman Ren, **Zhongkai Yu**, Yufei Ding, Adnan Aziz. “KPerfIR: Towards an Open and Compiler-centric Ecosystem for GPU Kernel Performance Tooling on Modern AI Workloads”
DAC’ 22	Zhuoran Song, **Zhongkai Yu**, Naifeng Jing, Xiaoyao Liang. “E2SR: an end-to-end video CODEC assisted system for super resolution acceleration”, Proceedings of the 59th ACM/IEEE Design Automation Conference, 2022
TCAD	Tianyun Ma, Yuanbo Wen, Xinkai Song, Pengwei Jin, Di Huang, Husheng Han, Ziyuan Nan, **Zhongkai Yu**, Shaohui Peng, Yongwei Zhao, Huaping Chen, Zidong Du, Xing Hu and Qi Guo. Harmonia: A Unified Architecture for Efficient Deep Symbolic Regression.

Internship

NVIDIA Research *Jun 2026 – Sep 2026*

Research Intern, Architecture Group, Manager: Po-An Tsai, Jenny Huang, and Christos Kozyrakis

Santa Clara, US

- Waferscale GPU architecture and system design for LLM

Samsung Semiconductor *Jun 2025 – Sep 2025*

Research Intern, AGI Lab, Manager: Shuyi Pei and Yangwook Kang *San Jose, US*

- MoE serving on multi-chiplet HBM-NMP architectures

Cambricon Technologies *Mar 2022 – Mar 2024*

Digital IC Design Intern, Architecture Group, Manager: Dong Han and Shaoli Liu *Beijing, China*

- Digital Front-end Design of a Commercial Cloud AI Chip Using Advanced Process
- Max-Power Test Cases for a Commercial AI Chip Using High-level Programming Language

Selected Projects

Patterns behind Chaos: Forecasting Data Movement for Efficient Large-Scale MoE LLM Inference

- Conducted a comprehensive data-movement-centric profiling of top tier MoE models (200B - 671B)
- Made systematic analysis and distrill six insights applicable for varied serving system design
- Proposed a future GPU architecture design to verify the insights

CLoRA: Cost-effective System for Multi-LoRA Model Inference using CXL

- Proposed a CXL-based near data processing architecture to support collaboration with GPU.
- Designed multiple collaborative approaches to utilize both GPU and NDP for multi-LoRA serving
- Built strategy selection and memory management system for optimal hardware utilization .

Cambricon-LLM: Accelerator for On-device Inference of 70B LLM

- Designed an innovative architecture to enable the inference of 70B-LLM on small edge devices.

- Reached a speed of 3.44 tokens for 70B LLM, and 22x to 45x speedup over existing technologies.

Collaborative Projects

KPerfIR: A Compiler-centric Profiler for GPU Kernel Performance Tooling

- Participated in building a compiler-centric profiler and tested it on NV/AMD GPUs.
- Used the profiler to improve naïve Triton FA-3 kernel

ReDM: CXL-based Memory Disaggregated System for Large-scale Recommendation Models

- Modified the original CXL memory extender architecture to support near data processing.
- Designed an access-aware strategy for offloading data to CXL memory and achieved a 2.37x speedup over traditional workload.

E2SR: An End-to-End CODEC Assisted System for Super Resolution (SR) Acceleration

- Enhanced SR algorithm acceleration on edge devices utilizing the computing power on cloud server.
- Developed E2SR/E2SR+ architecture on edge device to support end-to-end system integration.
- Proposed an adaptive framework responsive to varied environmental conditions.

Skills

Languages: Python, C, MATLAB, Verilog, CUDA, BANGC

Tools: Linux, (g)vim, Latex, VCS, Origin, Mathematica