Haoyang Zhang

zhang402@illinois.edu | hieronzhang.github.io | Haoyang Zhang
230 Coordinated Science Lab, 1308 W Main St, Urbana, IL 61801

Last updated: September 25, 2025

Research Interest

Computer architecture and system software. I'm interested in building efficient memory and storage systems/architecture for AI infrastructure, by exploiting algorithm-hardware co-design techniques.

Education

University of Illinois Urbana-Champaign 2022-2028 (exp.)
Ph.D., Computer Science
Advisor: Jian Huang

University of Michigan 2020-2022
(Dual) B.S.E., Computer Engineering

Shanghai Jiao Tong University 2018-2022
(Dual) B.S.E, Electrical and Computer Engineering

Publications [G]

Preprint

[1] Ziqi Yuan, **Haoyang Zhang**, Yirui Eric Zhou, Apoorve Mohan, I-Hsin Chung, Seetharami Seelam, and Jian Huang. "Cost-Efficient LLM Training with Lifetime-Aware Tensor Offloading via GPUDirect Storage". In: *arXiv preprint* (To Appear in NeurIPS 2025).

Conference Papers

[1] **Haoyang Zhang***, Yuqi Xue*, Yirui Eric Zhou, Shaobo Li, and Jian Huang. "SkyByte: Architecting an Efficient Memory-Semantic CXL-based SSD with OS and Hardware Co-design". In: *2025 IEEE International Symposium on High Performance Computer Architecture (HPCA)*. 2025.

[2] **Haoyang Zhang***, Yirui Zhou*, Yuqi Xue, Yiqi Liu, and Jian Huang. "G10: Enabling An Efficient Unified GPU Memory and Storage Architecture with Smart Tensor Migrations". In: *Proceedings of the 56th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*. 2023.

[3] Jiacheng Ma, Gefei Zuo, Kevin Loughlin, **Haoyang Zhang**, Andrew Quinn, and Baris Kasikci. "Debugging in the brave new world of reconfigurable hardware". In: *Proceedings of the 27th ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*. 2022.

[4] Xingyue Qian, Jian Shi, Li Shi, **Haoyang Zhang**, Lijian Bian, and Weikang Qian. "Scheduling Information Guided Efficient High-Level Synthesis Design Space Exploration". In: *2022 IEEE 40th International Conference on Computer Design (ICCD)*. 2022.

Impact of My Research

TeraIO: Cost-Efficient LLM Training with Lifetime-Aware Tensor Offloading via GPUDirect Storage

· We design and implement a new lifetime-aware tensor offloading framework for GPU memory expansion using low-cost PCIe-based solid-state drives (SSDs). It is developed explicitly for large language model (LLM) training with multiple GPUs and multiple SSDs.

· In comparison with state-of-the-art studies such as ZeRO-Offload and ZeRO-Infinity, TeraIO improves the training performance of various LLMs by 1.47x on average, and achieves 80.7% of the ideal performance assuming unlimited GPU memory.

SkyByte: Architecting An Efficient Memory-Semantic CXL-SSD with OS and Hardware Co-design

· The CXL-based solid-state drive (CXL-SSD) provides a promising approach towards scaling the main memory capacity at low cost. However, the CXL-based SSD has faced performance challenges due to the long flash access latency and unpredictable events such as garbage collection in the SSD device, stalling the host processor and wasting compute cycles. We present SkyByte, an efficient CXL-based SSD that employs a holistic approach to address the aforementioned challenges by co-designing the host operating system (OS) and CXL-SSD controller.

G10: Enabling An Efficient Unified GPU Memory and Storage Architecture with Smart Tensor Migrations

· We present a efficient unified GPU memory and storage architecture driven by the fact that DNN workloads are highly predictable. G10 integrates the host memory, GPU memory, and flash memory into a unified memory space, to scale the GPU memory capacity while enabling transparent data migrations. G10 utilizes compiler techniques to characterize the tensor behaviors in DNN workloads to schedule data migrations in advance by considering the available bandwidth of flash memory, host memory, and interconnections.

Industry Experience

T-head Division, Alibaba Cloud 2021
Software Research & Development Intern, *Host: Yunhai Shang*
• Implement the optimization for the jpeg library for RISC-V vector processors.

Selected Awards & Honors

MICRO 2023 Travel Grant 2023
Roger King Scholarship, University of Michigan 2021
University Honors, University of Michigan 2021
Dean's List, University of Michigan 2020, 2021

Teaching

University of Illinois Urbana Champaign

FA 2025	Graduate Teaching Assistant, Computer Architecture (CS 233)
SP 2025	Graduate Teaching Assistant, Computer Architecture (CS 233)
FA 2024	Graduate Teaching Assistant, Computer Architecture (CS 233)

Shanghai Jiao Tong University

SU 2022	Teaching Assistant, Computer Architecture (ECE 4700J / VE 470)
SU 2020	Teaching Assistant, Honors Physics (VP 160)

Academic Service

Artifact Evaluation Committee

HPCA 2025

Skills

Programming Languages

Proficient	C/C++, CUDA, C++ HLS, Verilog/System Verilog, Python
Familiar	Bash, Yacc, MATLAB, HTML
Capable	Tcl, Murphi

Frameworks/Technologies

ML Stack	PyTorch 2.7, FlashInfer
Simulators	gem5, MacSim, AccelSim (GPGPU-Sim), DRAMSim2
Profiling/Instrumentation	Intel PIN, Nvidia Nsight System, Nvidia Nsight Compute
ISAs	RISC-V, x86, ARMv8, MIPS
EDA Tools	Xilinx Vivado, Verilator, Synopsys VCS