Yutong Huang

yutonghuang.me
yutonghuang@ucsd.edu

EDUCATION

University of California, San Diego, La Jolla, CA Sep 2020 – Current

Pursuing PhD Degree in Computer Science (WukLab, advised by Prof. Yiying Zhang)

Master's degree in Computer Science (2023)

Purdue University, West Lafayette, IN Aug. 2014 – Dec. 2017

Bachelor of Science in Computer Engineering

Dean's List & Semester Honors (6 semesters, Spring 2015 – Fall 2017)

SKILLS

Coding Languages: C (User level and kernel level), C++, Python, Bash, CUDA, Verilog, Java

Frameworks/Tools: Linux, Pytorch, Nvidia Nsight, eBPF, LLVM, DynamoRIO, RDMA, Vivado

RESEARCH EXPERIENCE

LegoOS Distributed System (Best Paper)

Major Contributor Jun. 2017 – Sep. 2018

- Achieved a single user application running on distributed hardware without modification of the application by designing and implementing distributed user space virtual memory management using C language
- Reduced distributed system failure rate and increasing application memory access parallelism by designing memory replication using an asynchronous mechanism to hide performance overhead
- Resolved physical processors and memories resource allocation issue by designing a resource management system as a Linux kernel module
- Evaluated LegoOS system performance with Phoenix MapReduce, TensorFlow, and PARSEC Workload

Operating System for LLM Agent

Project Leader Oct 2025 – Current

- Analyzed LLM agent behavior across both CLI and AUC use cases, including full latency decomposition, RPC/call-count telemetry, and system-level performance profiling.
- Identified OS-level optimization opportunities by tracing redundant bBoN context spawns and proposing mechanisms for process and memory deduplication to reduce overhead.
- Prototyping a containerized execution model to run parallel agentic tasks with shared disk and memory layers (including AUC flows), enabling resource de-duplication without impacting user experience.

Linux based Far Memory Prefetch with ML

Project Leader Jan. 2023 – Sep 2025

- Proposed ideas of separating program context and runtime memory system context to greatly increase the predictability of memory access pattern
- Prototyped an LLVM compiler-based as proof of concept, achieve 30%-70% better performance than state of the art solutions
- Implemented the Linux kernel by modifying the kernel swap path, swap cache memory management system, and kernel swap eviction policies.
- Built an RDMA based kernel module as the far memory backend

Memory Efficient DNN Training *Project Leader* May. 2021 – Dec. 2022

- Identified the memory bottleneck of a DNN training jobs with FLOPS calculation and Nsight profiling
- Proposed an SGD variant to reduce memory footprint where backward pass activation data is approximated rather than stored in memory
- Integrated this SGD with pipeline parallelism to increase GPU utilization

Hardware-Software co-design system *Project Contributor* Sep. 2020 – Apr. 2021

- Designed user-friendly FPGA interface, enabling programmers with less hardware knowledge writing FPGA program under this memory system
- Implemented a simple pointer tracing example on FPGA memory system, reaching 100Gbps line rate
- Tested the simple program under this framework satisfying all the coherent assumptions provided by system

PUBLICATIONS

An Early Exploration of Deep-Learning-Driven Prefetching for Far Memory

Yutong Huang, Zhiyuan Guo and Yiying Zhang (NeurIPS 2025, MLsys Workshop)

LegoOS: A Disseminated, Distributed OS for Hardware Resource Disaggregation *(Best Paper Award)*

Yizhou Shan, Yutong Huang, Yilun Chen and Yiying Zhang (OSDI '18)

Clio: a hardware-software co-designed disaggregated memory system

Zhiyuan Guo, Yizhou Shan, Xuhao Luo, Yutong Huang, Yiying Zhang (ASPLOS '22)

Learned: Operating Systems

Yiying Zhang, Yutong Huang (ACM SIGOPS Operating Systems Review)

See the World Through Network Cameras

Yung-Hsiang Lu, George K Thiruvathukal, Ahmed S Kaseb, Kent Gauen, Damini Rijhwani, Ryan Dailey, Deeptanshu Malik, Yutong Huang, Sara Aghajanzadeh, Minghao Mina Guo (IEEE Computer 52)

WORK EXPERIENCE

Microsoft Research (Mentored by Sameh Elnikety) *Intern Researcher* May. 2019 – Aug. 2019

- Analyzed Latency behavior of heterogeneous virtual machine over-subscriptions in cloud environment. (standard VMs and credit based VMs)
- Designed micro-benchmark for evaluating the behavior of oversubscription of IO intensive workload and CPU intensive workload
- Evaluated disk IO latency of Hype-V hypervisors under oversubscription and compared results with KVM

AWARDS

Jacob School of Engineering Fellowship	Sep 2020
Jay Lepreau Best Paper Award at OSDI '18	Oct 2018
USENIX Student Grant for OSDI '18	Sep 2018