Luyi Li

lul014@ucsd.edu | Google Scholar Profile | LinkedIn Profile

Overview

I am a PhD candidate in the Department of Computer Science and Engineering at the University of California San Diego (UC San Diego), advised by Professor Dean Tullsen. My research interests span computer architecture and AI systems, including CPU microarchitecture, CXL memory systems, KV-cache management, and confidential computing for LLM serving. I enjoy understanding emerging systems, identifying their limitations, and designing practical solutions to improve performance, efficiency, and security.

Education

University of California, San Diego, La Jolla, CA, US — Sep 2022 – Jun 2027 (Expected)

- Ph.D. in Computer Science and Engineering
- Advisor: Prof. Dean Tullsen

Nanjing University, Nanjing, China — Sep 2018 – Jul 2022

- B.S. in Electrical Engineering
- Advisors: Prof. Lang Feng and Prof. Zhongfeng Wang

Work Experience

Research Intern at ByteDance, San Jose, CA, US — Jun 2026 – Sep 2026

- Topic: DPU-Managed KV Cache Offloading for Agentic LLM Serving
- Advisors: Dr. Misun Park and Dr. Wei Tang

Publications

[1] ***Luyi Li***, Nachuan Wang, Zhiye Song, Misun Park, Peng Xu, Wei Tang, Liguang Xie, Wu Xiang, Nam Sung Kim, Dean Tullsen, "**DPUKV: DPU-Managed KV Cache Offloading for LLM Serving**", submitted to **MLSys 2027**.

[2] ***Luyi Li****, Jangseon Park*, Nagendra Kumar Jamadagni, Zixuan Wang, Jishen Zhao, Tajana Rosing, Dean Tullsen (*Co-first authors), "**CXLeak: Characterizing and Exploiting Contention in Compute Express Link (CXL) Memory**", in **HPCA 2027**.

[3] Jangseon Park, ***Luyi Li***, Zixuan Wang, Junhee Park, Seunghak Lee, Jinin So, Jongryool Kim, Younghoon Min, Jishen Zhao, Dean Tullsen, Tajana Rosing, "**CXLens: From System Characterization to Validated Simulation of CXL-based Disaggregated Memory**", in **HPCA 2027**.

[4] Zixuan Wang, Suyash Mahar, ***Luyi Li***, Jangseon Park, Jinpyo Kim, Theodore Michailidis, Yue Pan, Tajana Rosing, Dean Tullsen, Steven Swanson, Kyung Chang Ryoo, Sungjoo Park, Jishen Zhao, "**The Hitchhiker's Guide to Programming and Optimizing CXL-Based Heterogeneous Systems**", in submission.

[5] ***Luyi Li****, Hosein Yavarzadeh*, Dean Tullsen (*Co-first authors), "**Indirector: High-Precision Branch Target Injection Attacks Exploiting the Indirect Branch Predictor**", in **USENIX Security 2024**.
Award: Distinguished Paper Award, IEEE Micro Top Picks 2025 Honorable Mention

[6] ***Luyi Li***, Jiayi Huang, Lang Feng, Zhongfeng Wang, "**PREFENDER: A Prefetching Defender against Cache Side Channel Attacks as A Pretender**", in **DATE 2022**.
Award: Best Paper Candidate

[7] Lang Feng*, Jiayi Huang*, ***Luyi Li***, Haochen Zhang, Zhongfeng Wang (*Co-first authors), "**RVDFI: A RISC-V Architecture with Security Enforcement by High Performance Complete Data-Flow Integrity**", in **Transactions on Computers**.

[8] ***Luyi Li***, Jun Lin, Zhongfeng Wang, "**PipeBSW: A Two-Stage Pipeline Structure for Banded Smith-Waterman Algorithm on FPGA**", in ISVLSI 2021.
Award: Best Paper Candidate

Selected Projects

Shared-GPU LLM Serving Across Confidential VMs Ongoing

- Profiled LLM inference in confidential VMs to identify and reduce KV cache offloading overhead.
- Implemented shared-GPU inference for multiple confidential VMs with shared model weights and an encrypted, isolated KV cache for each VM.
- Tools Used: vLLM, QEMU, Intel TDX, NVIDIA Confidential Computing

DPU-Managed KV Cache Offloading for Agentic LLM Serving Jun 2026 – Oct 2026

- Profiled vLLM and LMCache to identify memory-capacity and GPU-I/O bottlenecks in existing multi-tier KV-cache offloading systems.
- Designed and implemented DPUKV, which uses spare DPU compute and memory to manage the KV cache across GPU HBM, DPU DRAM, and remote storage without using host DRAM.
- Optimized DPUKV with a quantization kernel, pipelined transfers, and an agent-aware caching policy on the DPU, increasing KV transfer throughput while reducing offloading latency and remote storage and network traffic.
- Tools Used: vLLM, GPUDirect RDMA, GPUDirect Storage, NVMe-oF, NVIDIA BlueField DPU, DOCA SNAP

Profiling and Exploiting CXL Memory Expanders and Pools Apr 2024 – Dec 2025

- Designed the Heimdall benchmark to measure and profile the performance of modern cache-coherent heterogeneous systems.
- Evaluated the performance of various CXL expanders on both Intel and AMD platforms.
- Implemented a pointer-chasing microbenchmark to evaluate CPU cache utilization when using different memory devices.
- Reverse-engineered the physical address mapping of CXL memory and analyzed resource contention in CXL memory at different microarchitectural levels.
- Designed a high-throughput contention-based covert channel and side-channel attacks for website and LLM fingerprinting on the CXL memory controller.
- Proposed a software scheduler to reduce contention in CXL memory.
- Tools Used: C, x86 assembly, Python, kernel module development, llama.cpp

Reverse Engineering and Attacking Intel Indirect Branch Predictors Apr 2023 – Apr 2024

- Reverse-engineered the Indirect Branch Predictor (IBP) and Branch Target Buffer (BTB) microarchitectures in the latest Intel processors.
- Analyzed Intel's hardware defenses (IBPB, IBRS, STIBP), discovered previously unknown coverage gaps, and documented the precise conditions under which these mitigations fail.
- Developed high-precision Branch Target Injection (BTI) attacks against both IBP and BTB, and demonstrated how to leverage IBP/BTB states to break ASLR.
- Tools Used: x86 assembly, NASM, C, Python, Intel performance monitor
- Repository: github.com/owenlly/Indirector_Artifact (60+ stars)

Skills

Programming languages: C++, C, Python, Verilog, x86 / ARM / RISC-V Assembly

Platforms & tools: gem5, FPGA, llama.cpp, vLLM, SGLang

Technical Expertise: CPU and GPU Microarchitecture, CXL Memory, NVIDIA Confidential Computing, KV Cache Offloading, GPUDirect RDMA, GPUDirect Storage.

Services

Teaching assistant: CSE 12 - Basic Data Struct & OO Design, CSE 240A - Computer Architecture