Zesen Liu

igns.top | zesenliu@smail.nju.edu.cn | github.com/ftyghome | (+86) 152-5952-5167

Education

Nanjing University Sep. 2023 - Jun. 2026 (Expected)
M.S. in Computer Science and Technology | Institute of Computer Software
Advisor: Yanyan Jiang

South China University of Technology Sep. 2019 - Jun. 2023
B.S. in Computer Science and Technology | Embedded and Intelligent Robotics Laboratory GPA: 3.89/4
Advisor: Sheng Bi

Research Experience

PerfLab: Modeling Performance Characteristics in Mobile Systems 2024 - Present

Employs an experimental-data-driven approach to model system performance, aiming to uncover performance characteristics of applications in mobile systems.

- Introduces a "Lab" for performance observation, which is capable of constructing detailed performance profiles through rich datasets generated by experiments (e.g., adjusting resource allocation or injecting latency into specific operations).
- Utilizes statistical models and machine learning to extract performance characteristics and reason about the root cause of performance degradation.
- Allows developers to design new experiments easily. Developed a JIT-compatible, low-overhead hooking mechanism on Android, provides an eBPF-style API for flexible and unified instrumentation in JVM and kernel. Abstracts program execution into event sequences and enables developers to register hooks using PerfLab DSL to apply custom resource policies.
- Supports cross-application comparisons as well as intra-application analyses of thread- and function-level sensitivity to resource supply.

LockPerf: Hardware-Assisted Joint User-Kernel Space Tracing 2023 - 2024

Aligns OS scheduling events with user-space events on a single timeline, enabling low-overhead, "God's-eye view" analysis of application latency.

- Leveraged Intel Processor Trace (Intel PT) for hardware-assisted, high-precision event recording. Provides comprehensive performance reports by capturing all kernel scheduling events and specified user-space events (e.g., mutex lock/unlock for lock analysis) recorded via hardware instructions.
- Enabled detailed lock analysis: for each lock, extracts its initialization time, counts of fast and slow path entries, threads handovers, and timestamps of contention events. Pinpoints the exact moments when application latency is caused by suboptimal scheduling.

AdapCHOL: Sparse Cholesky Decomposition via HW-SW Co-design for Graph Optimization 2022 - 2023
Awarded Outstanding Undergraduate Thesis of South China University of Technology

A hardware-software co-designed solution for graph optimization algorithms. Addressed the challenges of limited computing resources in embedded systems, achieving a performance improvement of up to 148%.

- Implemented the Cholesky decomposition of sparse symmetric positive-definite matrices on the Xilinx Kria K26 SoM. Considering the hardware characteristics of FPGAs with unified memory from the outset, achieved a fully pipelined architecture for data loading, computation, and storage.
- Specifically optimized for small-scale matrices commonly encountered in robotic obstacle avoidance and navigation scenarios. Redesigned the dataflow model of Cholesky to reduce the latency of individual matrix decomposition operations.
- Developed efficient software scheduling and resource allocation algorithms for co-scheduling FPGA Compute Units (CUs) and the host CPU. Designed a software-level data consistency management framework based on AXI features to minimize unnecessary cache flush operations.

Projects & Awards

RadeonFlow: High-Performance DeepSeek Operators on AMD Instinct MI300X May. 2025

🏆 **Showcased by CEO Lisa Su at the AMD Advancing AI 2025 event, awarded $100,000 Grand Prize**

RadeonFlow implements FP8 Blockwise GEMM, MoE, and MLA operators, achieving up to 8x speedup compared to the official PyTorch baseline for MI300X.

- Implemented CCX-aware scheduling for the AMD CDNA 3 architecture to reduce inter-CCX communication overhead in GEMM kernel.
- Bypassed ROCm compiler and utilized manual Async Copy optimization and instruction interleaving to efficiently exploit the GPU hardware pipeline.
- Eliminates CPU synchronization overhead by employing indirect pointers when invoking the hipBLAS library.

TinyBACC: A High-Performance Compiler based on LLVM IR Aug. 2021

TinyBACC is a LLVM IR-based compiler focused on simplicity and performance. Its generated binaries achieve performance comparable to GCC -O3. This project won the National Second Prize in the National Collegiate Computer System & Programming Contest - Compiler Design Track 2021.

- Features a complete front-end and back-end design (C→LLVM IR→ARM Thumb), supporting the integration and unified management of various optimization passes.
- Includes optimization passes such as GVN/GCM, Loop Strength Reduction, and pattern-based auto-vectorization. Manually tuned instruction scheduling strategies based on A53 core hardware features to efficiently utilize the processor pipeline.

DomCast: Remote UI Rendering for Mobile Devices Sep. 2024

DomCast is the first low-overhead screencasting solution that proposes UI Tree projection and remote rendering. It offloads rendering tasks to the receiver, transmitting only compressed UI component properties. This reduces data transmission to 8% of the original while achieving a high frame rate of 60 FPS. Deployed on OpenHarmony 5.0, this project won the First Prize in the 1st China Open Source Innovation Competition on Operating Systems.

ACM/ICPC Programming Contests 2020 - 2022

- Silver Medal, The 46th ICPC Asia Regional Contest
- Bronze Medal, 2021 China Collegiate Programming Contest (CCPC), Weihai Site
- Bronze Medal, 2020 China Collegiate Programming Contest (CCPC), Mianyang Site

Technical Skills

Compilers **Proficient in implementing and tuning high-performance CPU compilers**

- Familiar with the entire workflow from code parsing to machine code generation.
- Understand the design and engineering of several LLVM passes, capable of debugging and root-causing compiler bugs; proficient with LLVM IR and able to develop new passes for applications like kernel driver memory access pattern analysis.
- Experience in implementing a compiler for a FP language (MoonBit).

GPU/FPGA **Experienced in designing acceleration algorithms for heterogeneous platforms**

- (**GPU**) Proficient in operator design and tuning, understand the hardware characteristics for modern data center GPUs; familiar with asynchronous memory access implementation and compiler support, with experience in hacking GPU compilers.
- (**FPGA**) Programming experience with C++ variants for FPGA architectures (Xilinx Vitis High-Level Synthesis), skilled in designing acceleration algorithms to maximize pipeline utilization and in HLS-based performance tuning and HW-SW co-design.

Embedded & Others Proficient in chip soldering and microcontroller development; familiar with robot obstacle avoidance and navigation algorithms and related hardware-software co-acceleration solutions; holder of a Class A Amateur Radio Operation Certificate.

Languages **IELTS 7.5** (Speaking 6.5)

Community Contributions & Activities

Compilers & Systems Contributed to LLVM's Interprocedural Sparse Conditional Constant Propagation (IPSCCP) pass (authored a technical blog post (in Chinese), submitted a NFC patch, currently refactoring the pass to enable more optimization opportunities on glibc). Added Boot from NAND support to the RISC-V Allwinner-D1 SPL (code merged into the mainline U-Boot Bootloader). Added external-rootfs support to podman-compose. Creator of pwprof, a memory access profiler. More projects available on GitHub.

Embedded & Robotics Open-source contributions include: the Unity-ROS simulation library siemens/ros-sharp ; the Webots simulation framework for ROS2; porting of the laser-line-extraction library for laser line detection.

ROCm/Xilinx Community Presented "From Assembly to PyTorch: A Full-Stack Design and Tuning of DeepSeek Operators on MI300X" at the AMD Developer Meetup 2025. Published a locally deployed spreadsheet AI agent on Hackster: PowerSheet: Powerful AI Synthesizer for Spreadsheets based on AMD Ryzen NPU.

Teaching & Activities

- TA for "Operating Systems: Design and Implementation" at Nanjing University for the 2024 and 2025 academic years.
- Media Chair for ACM SIGOPS ChinaSys'25.

Last Updated on August 2025