ZHENGXIONG LI

608-658-4205 | zhengxiong.li@wisc.edu | zhengxiongli08.github.io | linkedin.com/in/zhengxiong-li

EDUCATION

University of Wisconsin-Madison — Madison, USA
Ph.D., Department of Electrical and Computer Engineering — *09/2024 – Present*

Wuhan University — Wuhan, China
B.S., Microelectronics Science and Engineering, Hongyi Honor College — *09/2020 – 06/2024*

PUBLICATIONS

- **Li, Z.**, Huang, T., Ogras, U. (2026). *SET: Stream-Event-Triggered Scheduling for Efficient CUDA Graph Pipelines*. International European Conference on Parallel and Distributed Processing (Euro-Par), Pisa, Italy.
- Jie, T., **Li, Z.**, Ogras, U., Huang, T. (2025). *A Scalable Code Generation Flow for Heterogeneous Parallel RTL Simulation using MLIR*. IEEE High Performance Extreme Computing Conference (HPEC). [DOI]
- Li, X.†, **Li, Z.**†, et al. (2024). *MSGM: An Advanced Deep Multi-Size Guiding Matching Network for Whole Slide Histopathology Images Addressing Staining Variation and Low Visibility Challenges*. IEEE Journal of Biomedical and Health Informatics (JBHI). [DOI] †Equal contribution.
- Li, X.†, Hu, T.†, **Li, Z.**†, et al. (2024). *MRRM: Advanced Biomarker Alignment in Multi-Staining Pathology Images via Multi-Scale Ring Rotation-Invariant Matching*. IEEE Journal of Biomedical and Health Informatics (JBHI). [DOI] †Equal contribution.
- Hu, M., Jiang, X., **Li, Z.**. (2023). *A Resource-efficient FIR Filter Design Based on an RAG Improved Algorithm*. International Conference on Circuits and Systems (ICCS). [DOI]

RESEARCH EXPERIENCE

SET: Stream-Event-Triggered Scheduling for Efficient CUDA Graph Pipelines — Madison, USA
Research Assistant; Supervisor: Prof. Tsung-Wei Huang and Prof. Umit Ogras — *06/2025 – 11/2025*

- Developed a high-performance runtime framework (SET) for CUDA graphs to optimize GPU workloads scheduling; decomposed monolithic processing graphs into asynchronous, multi-stage pipelines to achieve overlap of memory transfers and computation.
- Designed a dynamic work-stealing scheduler to handle load balancing across multiple CUDA streams; utilized event-driven synchronization and non-blocking CUDA callbacks to decouple host scheduling overhead from device execution.
- Implemented a memory-safe, just-in-time parameter update mechanism for persistent CUDA graph executables, enabling the dynamic reuse of graph handles for different data payloads without the cost of re-instantiation.
- Conducted extensive performance analysis using NVIDIA Nsight Systems and Nsight Compute to identify synchronization bottlenecks; optimized pipeline depth and resource allocation, achieving up to 44% higher throughput and reducing scheduling overheads by up to 54% compared to baselines.

PROWESS: Processor Reconfiguration for Wideband Spectrum Sensing — Madison, USA
Research Assistant; Supervisor: Prof. Umit Ogras — *09/2024 – 03/2025*

- Implementing a system-level domain-specific SoC simulation framework (DS3) on Zynq platform. Porting the core on-chip scheduler module from Python to C to meet resource restrictions on embedded systems. Finding the optimal resource configuration that can guide the hardware team's design.
- Studied the communication overheads between on-chip scheduler (PS side) and program launcher (PL side) through AXI protocols. Found the optimal job distribution for task scheduling to achieve best efficiency and performance.
- Optimized the DS3 simulator to enable massive parallelization, including decoupling improper program design and refactoring stale data structures. Achieved up to 5x speedup.
- Performed design space exploration based on DS3 simulator and analyzed the generated data. Studied the impact on throughput caused by different factors, including processing elements (PEs) mesh architecture, iMEM size, communication overheads, etc. Identified the bottleneck to achieving higher throughput gains.

Development of an FPGA-Based Automatic Calibration System for RFIC — Wuhan, China
Undergraduate Researcher; Supervisor: Prof. Xianyang Jiang — *05/2022 – 04/2023*

- Designed a pair of SPI protocol communication interfaces to enable data transmission among the PC, FPGA board, and target chip to be calibrated.

- Employed the DDS method to develop a system generating periodically vibrating waveforms with high stability.
- Developed a series of FIR filters and verified their efficacy using Matlab while integrating various decimation modules to filter the digital data acquired from ADC for precise signal processing.
- Conducted simulation and debugging procedures for FPGA and PCB boards, leading to their successful deployment.

FPGA/GPU-Based Numerical Solvers for Sustainable Energy System Simulation Edmonton, Canada
Research Intern; Mitacs Internship; Supervisor: Prof. Venkata Dinavahi *06/2023 – 09/2023*

- Developed a high-performance solution for real-time simulation of electrical systems, ensuring the hardware's ability to predict the next 1us' situation within a strict time frame, such as 100ns or shorter.
- Trained GRU and LSTM models for Synchronous Machine and Induction Machine modeling.
- Implemented the pre-trained model on Xilinx Data Center Acceleration Card (U250), aiming to achieve real-time or faster than real-time simulation for the entire electrical system.

Multi-Scale Spatial Registration Method for Whole-Slide Pathology Images Wuhan, China
Undergraduate Researcher; Supervisor: Prof. Cheng Lei *03/2023 – 05/2024*

- Reviewed existing literature on pathology image analysis and identified limitations in current methods such as SIFT, KAZE, AKAZE, and BRISK.
- Developed an algorithm based on multi-scale spatial space for efficient extraction of image features; overcame challenges posed by rotation, missing parts, and color differences in whole-slide pathology image analysis.
- Successfully achieved image registration based on the extracted features and achieved a high level of accuracy.

Coursework Projects

GPU Benchmarks Porting and Simulation for AMD GPUs on gem5 Madison, USA
Project Leader *09/2024 – 12/2024*

- Ported the Polybench and Lonestar benchmark suites from CUDA to AMD HIP, manually resolving inline assembly and library mismatches to enable cross-platform compatibility.
- Simulated AMD Vega 20 microarchitecture using the gem5 simulator to analyze performance characteristics, validating simulation accuracy against real-world Radeon VII hardware execution.
- Investigated the impact of dynamic vs. static register allocation schemes on irregular graph algorithms (BFS, SSSP), identifying resource contention bottlenecks and potential logic errors within the simulator's scheduling model.

RISC-V CPU Implementation on FPGA Wuhan, China
Core Member *08/2022 – 02/2023*

- Designed a five-stage pipeline CPU and achieved basic operations such as XOR, AND, and algorithm calculation.
- Designed bypass and stall structure to solve data hazard while data are required before it is written back into registers.

FPGA-Based Ultra-High-Speed Panoramic Camera Image Stitching Project Wuhan, China
Core Member *04/2023 – 06/2023*

- Improved the performance of existing software algorithm and transformed it into hardware description language.
- Implemented the improved algorithm on a Xilinx FPGA development board, to achieve high resolution and high frame rate for panoramic camera image stitching.
- Collaborated with a diverse team, including fellow students from the laboratory, engineers from the company, and students from Huazhong University of Science and Technology.

Professional Service

Teaching Assistant of Digital Logic Circuit Laboratory Experiment Course 04/2023 – 06/2023

- Conducted seminars, collected technical documents, graded them, and contributed to curriculum design.
- Successfully recruited exceptional students to the laboratory through targeted outreach efforts.

Skills & Tools

Programming: CUDA, C & C++, Python, LLVM, MLIR, Verilog, Matlab
Laboratory Techniques: Nsight Systems Profiler, GDB, Vitis, Vivado, Quartus, Modelsim, FPGA hardware debug, FPGA software development
Languages: Mandarin (Native), English (Fluent)