Zhiyu Wu

Zhiyuwu2@illinois.edu • https://xzzwzy.github.io/ • 7345968166 • 410 N Lincoln Ave

EDUCATION

University of Illinois Urbana-Champaign **Champaign, IL**
M.S. Computer Science – Thesis Based
Research Interest: LLMs, ML system infrastructure, LLM systems *Expected May 2026*

University of Michigan **Ann Arbor, MI**
B.S.E. Computer Engineering (Dual Degree) *August 2022 – May 2024*
GPA: 3.58 / 4.00

Shanghai Jiao Tong University **Shanghai, China**
B.S.E. Electrical and Computer Engineering (Dual Degree) *Sept. 2020 – August 2024*
GPA: 3.75 / 4.00

Research Interest: LLMs, ML system infrastructure, systems for LLMs
Coursework: Computer Architecture, Operating Systems, Computer Network, Machine Learning, Embedded Systems

RESEARCH EXPERIENCE

Research Assistant in GAEA Lab **Champaign, IL**
Supervisor: Fan Lai *July 2024 - present*

- Classify LLM serving requests into three categories based on unique system objectives:
  - Latency-Intensive: For streaming use case, ensuring fluent reading experience.
  - Throughput-Intensive: Only focus on the job completion time (JCT).
  - Bulk Requests: Large groups of requests submitted together, with collective completion time as the priority.
- Define Service Level Objectives (SLO) for each request type:
  - Latency-Intensive: SLO based on Quality of Experience (QoE).
  - Throughput-Intensive: Extended deadlines, calculated as the time it runs alone on the machine multiplied by a scaling ratio.
  - Bulk Requests: SLO based on the deadline of the last request in the group.
- Develop an SLO-aware scheduling policy using length prediction to optimize job completion time (JCT) and improve user experience in LLM inference.
  - The policy combines DAG scheduling and two-dimensional knapsack scheduling, ensuring efficient resource allocation to meet SLOs across different request types.

Research Assistant in Symbiotic Lab **Ann Arbor, MI**
Supervisor: Mosharaf Chowdhury *May 2023 – April 2024*

- Identified that in LLM text-streaming services, systems must generate faster than user reading speed to enhance user experience, addressing gaps in previous metrics.
- Defined Quality of Experience (QoE) in LLM serving by tracking each step of text generation and monitoring the overall user experience throughout the entire streaming process.
- Formulated the problem as a knapsack optimization and developed a scheduling algorithm to maximize QoE in online LLM serving.
- Built Andes, an LLM serving system on top of vLLM, integrating the scheduling algorithm to enhance QoE in real-time LLM services.
- Co-authored the paper "Andes: Defining and Enhancing Quality-of-Experience in LLM-Based Text Streaming Services" as the second author.

PUBLICATIONS

- Andes: Defining and Enhancing Quality-of-Experience in LLM-Based Text Streaming Services; Preprint, 2024; Jiachen Liu, **Zhiyu Wu**, Jae-Won Chung, Fan Lai, Myungjin Lee, Mosharaf Chowdhury

PROJECT EXPERIENCE

Symbiotic Lab/ML.ENERGY.LEADERBOARD Team **Ann Arbor, MI**
Developer *May 2023 – Sept. 2023*

- Developed the ML.ENERGY Leaderboard, an open-source platform for benchmarking the energy efficiency and NLP performance of LLM models.
- Defined performance metrics and implemented scripts for optimized batched inference to ensure accurate measurement.
- Contributed to the online Chatbot Arena which gathers data on models' energy consumption and performance.

Toy Operating System **Ann Arbor, MI**

- Created a toy operating system with physical memory and disk management.
- Implemented read-write locks using mutexes to manage multi-threading.
- Developed virtual memory management with a page table and a network file server using sockets.
- Built a custom file system for networked access.

Out-of-order Execution Pipeline for the MIPS R10K Microprocessor **Shanghai, China**

- Developed an out-of-order execution pipeline with six stages on the MIPS R10K microprocessor.
- Implemented key components including register renaming, reservation station, reorder buffer, and a common data bus for enhanced parallelism.
- Applied Tomasulo's algorithm for dynamic scheduling and reducing pipeline stalls.
- Added a Load Store Queue and a Branch Target Buffer to further optimize execution efficiency and improve instruction throughput.

Video Streaming via CDN **Ann Arbor, MI**

- Developed a proxy server for handling video streaming across multiple clients and servers, ensuring scalability and reliability.
- Implemented adaptive bitrate streaming to minimize buffering and enhance user experience based on real-time network conditions.
- Used DNS load balancing with round-robin and distance-based server selection, utilizing Dijkstra's algorithm to optimize server choices based on proximity and load.

Static Router **Ann Arbor, MI**

- Built a static router with basic packet forwarding capabilities to route real packets to HTTP servers.
- Implemented layer 2 and layer 3 protocols, including ARP, ICMP, and Ethernet, for routing and handling network traffic.

Embedded Device for Keystroke Timing and Acoustic Attack Protection **Shanghai, China**

- Designed the device to intercept keystrokes and introduce random delays before sending to the PC.
- Implemented keystroke sound playback to counter acoustic attacks using recorded sounds.
- Based the system on the STM32F405 microcontroller with Embedded Rust for secure and efficient performance.
- Delivered a compact, user-friendly design with production costs around $24.50 per unit.
- Utilized SD card for storing custom keystroke sounds and USB peripherals for communication with keyboard and host PC.

PROFESSIONAL SERVICE

- VP 160 Honors Physics SJTU, 2022 Summer

SKILLS

***Computer**:* C++, C, Python, Rust, Pytorch, CUDA, System Verilog, Embedded C/Rust, Linux, MATLAB, Git, LaTeX

HONORS

Dean List, *Umich* 2023
University Honor, *Umich* *2022*
Tang Junyuan Scholarship, *SJTU* *2022*
SJTU Undergraduate Excellent Scholarship Class B, *SJTU* *2022*