SONGTING (MICHAEL) WANG

stw183164761@gmail.com ◇ linkedin.com/in/songting-wang ◇ +1 (412) 916-4409

EDUCATION

Carnegie Mellon University — **Pittsburgh, PA**
M.S. in Electrical and Computer Engineering — **August 2025 - May 2026**
B.S. in Electrical and Computer Engineering, Minor in Computer Science — **August 2022 - Dec 2025**
- **Coursework**: Distributed Systems, AI/ML Systems, Deep Learning Systems, CPU/GPU Architecture, Algorithms, Functional Programming

WORK EXPERIENCE

Software Engineer Intern - Core Infra — **May 2025 - August 2025**
S&P Global - Kensho, Boston, MA

- Implemented **structured logging** for OpenSearch log-router, enabling field queries and reducing log analytics time by **50%**.
- Delivered **endpoint-level** FastAPI metrics for **7+ product teams** using Grafana and Prometheus.
- Built and managed a centralized, automated **alert** visualization system on **Kubernetes** for **60+** application teams.

Software Engineer - Backend — **October 2023 - May 2024**
EnterviewAI, Pittsburgh, PA

- Led a **10-member** team to deliver production-grade backend services (e.g., Recording, Question Bank) using **Django** and **Azure**.
- Implemented AI features (e.g., Study Plans, Brainstorming, Mock Interviews, Feedback) with **AutoGen** and **LangChain**.

Software Engineer Intern - ML Infra — **May 2024 - August 2024**
ZKH Industrial Supply, Shanghai, China

- Developed an **LLM Benchmark System** using **Flask** and **Streamlit**, reducing training evaluation time by **25%**.

Software Engineer Intern - Backend — **Dec 2023 - Jan 2024**
United Imaging Intelligence, Shanghai, China

- Enhanced ShanghaiTech University's **medical questionnaire app**, extending question abstractions and backend update semantics.

RESEARCH EXPERIENCE

Graduate Researcher *with Google CoreML × CMU Catalyst* — **Feb 2026 – Present**
Google, Pittsburgh, PA

- Building a **compiler** to lower ***Mirage***-generated computation graphs into efficient Google **TPU kernels** (advised by *Prof. Zhihao Jia*).

ML Systems Research *with Prof. Zhihao Jia* — **August 2025 - Present**
CMU Catalyst Group, Pittsburgh, PA

- Working on ***Mirage Persistent Kernel*** (arXiv Link), a Compiler and Runtime for Mega-Kernelizing Tensor Programs.
- Implemented FlashInfer's optimized **Gumbel-Max Sampling** CUDA kernels in Mirage Persistent Kernel and verified with unit tests.
- Working on **Expert-Parallel Mixture-of-Experts** CUDA kernels in Mirage Persistent Kernel using all-to-all (A2A) communication.

Data Systems Research *with Prof. Vyas Sekar* — **September 2024 - Present**
CyLab Security & Privacy Institute, Pittsburgh, PA

- Invited **speaker** at ***Current 2025***, the world's largest Data Streaming conference attended by **3,500+** industry professionals.
- Open-Sourced ***FlinkSketch***, a high-performance library of sketching algorithms for Apache Flink.
- Building ***ProjectASAP***: Low-latency, cost-efficient data pipelines to support next-gen agentic workloads.

TEACHING EXPERIENCE

Teaching Assistant — **January 2024 - May 2025**
Carnegie Mellon University, Pittsburgh, PA

- **Distributed Systems (15-440/640)**: Supported 200 students on distributed protocols; led recitations; designed/graded coursework.
- **Computer Systems (18-213/613)**: Supported 150 students on assembly, memory, I/O, concurrency; led small-groups; graded work.

PROJECT EXPERIENCE

Needle Deep Learning Framework – C++, Python — **Sept 2025 - Dec 2025**

- Built core components of Needle, including **autodiff, tensor IR, backend dispatch, common neural network modules**.
- Implemented **FlashAttention** in **CUDA kernels**, improving memory and throughput efficiency.

Distributed File System – C, Java — **January 2024 - May 2024**

- Implemented **Remote Procedure Call (RPC)** and **concurrent LRU** caching to support efficient client-server communication.
- Achieved reliability and scalability through **two-phase commit**, **write-ahead logging**, and **dynamic scaling**.

RISC-V CPU Microarchitecture – C, Verilog — **January 2025 - May 2025**

- Implemented a 7-stage pipelined **RISC-V** CPU microarchitecture, with **LRU Cache** and optimizations to improve perf/watt.

SKILLS

Languages	Python, C++, Java, C, Jsonnet, Scala, JavaScript, Standard ML, C#, Verilog, Rust
Frameworks	CUDA, JAX Pallas, Kubernetes, Flink, PyTorch, Numpy, Django, Flask, FastAPI, React, Node.js, Streamlit
Tools	Git, Grafana, Prometheus, Terraform, Jenkins, OpenSearch, Azure, AWS, Docker, Postman, Databases