SONGTING (MICHAEL) WANG

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

EDUCATION

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

WORK EXPERIENCE

Machine Learning Engineer, Annapurna Labs — **July 2026 - Present**
Amazon Web Services, Seattle, WA

- **Inference for custom AI accelerators**

Software Engineer Intern — **May 2025 - Aug 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 — **Oct 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 — **May 2024 - Aug 2024**
ZKH Industrial Supply, Shanghai, China

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

RESEARCH

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

- Building a **transpiler** to lower ***Mirage***-generated computation graphs into efficient **Google TPU kernels** (advised by *Prof. Zhihao Jia*).
- Implemented **JAX/Pallas kernels** for core LLM operators, including **GQA attention** and **RoPE**.
- Designed a **semantic backend IR** to decouple op semantics from codegen, enabling modular kernel selection for different hardware.

ML Systems Research *with Prof. Zhihao Jia* — **Aug 2025 - May 2026**
CMU Catalyst Group, Pittsburgh, PA

- Working on ***Mirage Persistent Kernel***, a Compiler and Runtime for Mega-Kernelizing Tensor Programs.
- Worked on **Expert-Parallel Mixture-of-Experts** and FlashInfer's **Gumbel-Max Sampling** CUDA kernels.
- Implemented **Hash Routing** gate support for the **DeepSeek V4** model.

Data Systems Research *with Prof. Vyas Sekar* — **Sept 2024 - April 2025**
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.

PUBLICATION

Mirage Persistent Kernel: A Compiler and Runtime for Mega-Kernelizing Tensor Programs — **OSDI '26**

TEACHING

Teaching Assistant — **Jan 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.

PROJECTS

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

- Built core Needle components (**autodiff, tensor IR, backend dispatch, NN modules**) and implemented **FlashAttention** with **CUDA**.

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

- Built systems using **RPC**, **concurrent LRU caching**, **two-phase commit**, **write-ahead logging**, and **dynamic scaling**.

RISC-V CPU Microarchitecture – C, Verilog — **Jan 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
Technologies CUDA, JAX, PyTorch, Flink, Kubernetes, Docker, Grafana, Prometheus, OpenSearch, Django, Flask, FastAPI, Git, Databases