Shanli Xing

Website: xsl.ing
Email: shanlx@cs.washington.edu

Research Interests

My research interests lie in building efficient AI systems and creating AI-driven workflows that continuously optimize them.

Education

University of Washington Seattle, WA
Bachelor of Science in Computer Science with Honors *Sep 2022 – Jun 2026*
GPA: 3.74/4.00;
Advised by Prof Luis Ceze

Publications & Articles

* Equal Contribution

[P1] FlashInfer-Bench: Building the Virtuous Cycle for AI-driven LLM Systems
Shanli Xing*, Yiyan Zhai*, Alexander Jiang*, Yixin Dong*, Yong Wu, Zihao Ye, Charlie Ruan, Yingyi Huang, Yineng Zhang, Liangsheng Yin, Aksara Bayyapu, Luis Ceze, Tianqi Chen
In Submission to MLSys 2026

[A1] Sorting-Free GPU Kernels for LLM Sampling
Shanli Xing, Zihao Ye, Bohan Hou, Luis Ceze, Tianqi Chen
Technical Blog Post

Research Experiences

Catalyst, Carnegie Mellon University Pittsburgh, PA
Visiting Undergraduate Researcher, advised by Prof. Tianqi Chen *Jun 2025 - Oct 2025*

- **FlashInfer-Bench: Building the Virtuous Cycle for AI-driven LLM Systems [P1]**
  - Co-led the FlashInfer-Bench project, an infrastructure providing a robust evaluation environment and rapid deployment path for AI-generated GPU kernels; coordinated a core team of four for fast project scaffolding and internal iteration.
  - Designed the FlashInfer Trace schema. Specified kernel interfaces and wrote reference implementations for the initial kernel dataset (including attention, GEMM, sampling, and fused MoE operators).
  - Led the architecture design and primarily implemented the Python library, including the distributed benchmarking engine, the apply module (kernel deployment feature), integration with FlashInfer and serving engines, the kernel build pipeline, and the command-line interface. Built and deployed the web leaderboard.
  - Co-authored the release blog post and the accompanying research paper.

SAMPL, UW CSE Seattle, WA
Undergraduate Researcher, advised by Prof. Luis Ceze, mentored by Dr. Zihao Ye *Jan 2024 - Jun 2025*

- **FlashInfer: Kernel Library for LLM Serving [Github][A1]**
  - Contributed to FlashInfer, a high-performance GPU kernel library for LLM serving.
  - Co-designed a dual-pivot rejection sampling/renormalization algorithm and implemented corresponding GPU kernels. Developed the `LogitsProcessor` module, softmax kernel, SM90 cutlass grouped-GEMM kernel and related Python interfaces.
  - Authored the technical blog post *Sorting-Free GPU Kernels for LLM Sampling*, demonstrating how rejection sampling can eliminate the top-k/top-p sorting bottleneck and achieve over 50% speedup compared to the PyTorch baseline.

Talks & Media Coverages

FlashInfer-Bench [P1] *2025*

- Gave in-person talk at CMU Catalyst.
- Project release featured by NVIDIA AI Developer, vLLM, and LMSYS Org (SGLang) .

Personal Projects

Radicle Github
Course Project of Systems for ML

- A demo inference engine implemented with Ray, featuring prefill/decode disaggregation, runtime elasticity, node multiplexing, continuous batching, and token streaming.

ChatFeedback Github
Course Project of Data-centric ML

- A full-stack web application to sustainably crowdsource high-quality, scalable, and diverse human preference data for LLM alignment.

Trefoil
Course Project of Programming Languages

- A Lisp-like language implemented in OCaml with an S-expression frontend and an LLVM backend that compiles to LLVM IR and native executables.

ReedIsland Github
Personal Android app project

- An Android forum client. Reached 1,000+ active users according to Visual Studio App Center.

Miscellaneous

Awards & Honors

- UW Annual Dean's List (2022-2025)

Selected Coursework

- **Graduate:** Systems for ML, Distributed Systems, Deep Learning, Natural Language Processing, Advanced Machine Learning, Data-centric Machine Learning, Ethics in AI
- **Undergraduate:** Computer Vision, Computer Graphics, Systems Programming, Programming Languages, Software Design & Implementation, Linear Algebra, Differential Equations

Technical Skills

- Programming: Python, C/C++, Rust, OCaml, CUDA, Triton, Java/Kotlin, Typescript
- Tools: Git, Docker, CMake, Gradle, NVIDIA Nsight
- Frameworks: PyTorch, Ray, vLLM, SGLang, Next.js, Android SDK