Yuxing Xiang

Peking University, No.5 Yiheyuan Road, Beijing, China

Email: xiangyx@stu.pku.edu.cn

GitHub: EchoStone1101

EDUCATION

- **Peking University** — Advisor: Xin Jin
  Ph.D. in Computer Science — *Sep 2024 – Present*
- **Peking University** — GPA: 3.8/4.0 (top 5%)
  B.S. in Computer Science — *Sep 2020 – July 2024*

RESEARCH INTEREST

My research interests include system and algorithm design for generative AI (primarily large language models, LLMs), and practical formal verification. I am also exploring their potential crossover — *e.g.*, formally verified code generation via reinforcement learning (RL).

PUBLICATIONS

- **Iceberg: Automated Verification of DNS Authoritative Engines via Just-in-Time Summarization**
  Yuxing Xiang, Rilin Huang, Naiqian Zheng, Xin Jin
  USENIX Symposium on Networked Systems Design and Implementation (NSDI 2026)
- **ServeGen: Workload Characterization and Generation of Large Language Model Serving in Production**
  Yuxing Xiang, Xue Li, Kun Qian, Yan Zhang, Wenyuan Yu, Ennan Zhai, Xin Jin, Jingren Zhou
  USENIX Symposium on Networked Systems Design and Implementation (NSDI 2026)
- **Aegaeon: Effective GPU Pooling for Concurrent LLM Serving on the Market**
  Yuxing Xiang, Xue Li, Kun Qian, Yufan Yang, Diwen Zhu, Wenyuan Yu, Ennan Zhai, Xuanzhe Liu, Xin Jin, Jingren Zhou
  ACM Symposium on Operating Systems Principles (SOSP 2025)
- **Automated Verification of an In-Production DNS Authoritative Engine**
  Naiqian Zheng*, Mengqi Liu*, **Yuxing Xiang**, Ennan Zhai, Linjian Song, Dong Li, Feng Han, Nan Wang, Yong Ma, Zhuo Liang, Dennis Cai, Xuanzhe Liu, Xin Jin (*Equal Contribution)
  ACM Symposium on Operating Systems Principles (SOSP 2023)

EXPERIENCE

- **Xiaomi LLM Core** — Nov 2025 – Present
  Research Intern — *Mentor: Liang Zhao*
- **Tencent Wechat Group** — June 2025 – Nov 2025
  Research Intern — *Mentor: Hao Cao*
- **Alibaba Cloud** — May 2024 – June 2025
  Research Intern — *Mentor: Kun Qian*
- **Alibaba Cloud** — Feb 2023 – Sep 2023
  Research Intern — *Mentor: Mengqi Liu*

PROJECTS

Large Language Model Serving

- **Aegaeon** — **Dec 2024**
  With the rising popularity of Model-as-a-Service (MaaS) for LLMs, service providers are wasting precious GPUs on serving the "long-tailed" models with sporadic invocations, calling for an GPU pooling approach. Aegaeon is a multi-LLM serving system that performs model auto-scaling at the *token* granularity to enable truly effective GPU pooling – up to 7 models per GPU, whereas existing solutions remain limited to 2 or 3 due

to GPU memory capacity (multiplexing) or active model count of the workloads (request-level auto-scaling). Aegaeon achieves this through token-level scheduling and full-stacked optimizations of the preemptive auto-scaling sequence. Experiments show that Aegaeon sustains 2-2.5× higher request arrival rates or 1.5-9× more goodput compared to existing solutions. In its beta-deployment at Alibaba Cloud Model Studio, Aegaeon currently serves 47 models, reducing the number of GPUs required from 1,192 to 213 (82% saving).

- **ServeGen** **March 2025**
  ServeGen is a workload generation framework that generates realistic LLM serving workloads based on extensive characterization of large real production traces at Alibaba Cloud Model Studio, spanning across 3 categories (language, multimodal, and reasoning), 12 LLMs, and 4 months. Its core insight is *per-client causal modeling* – the complex shifting patterns in an aggregate workload can be captured by the rate fluctuation of several top clients, each with relatively predictable behavior. ServeGen aims to advance the status quo of LLM inference system benchmarking, which, represented by the long-time favorite "ShareGPT + Poisson" approach, is shown to yield misleading conclusions in our experiments.

Formal Verification

- **DNSV** **April 2023**
  DNSV is an automated verification framework designed for the in-house DNS engine at Alibaba Cloud, whose correctness is crucial to the Internet. To handle over 2,000 lines of weakly encapsulated, fast iterating production Golang code, DNSV adopts refinement proof based on a hybrid of automated code summaries (with symbolic execution over LLVM IR) and manual specifications (for core low-level operations). DNSV finds 9 bugs across multiple versions of the engine, with an porting effort of less than one person-week each.

- **Iceberg** **Jan 2024**
  Iceberg is an automated verification framework for DNS authoritative engines, that achieves low manual effort while being generalizable to multiple implementations. The core of Iceberg is *Just-in-Time Summarization*, which scales automated refinement proof for *all* code modules by conducting summarization only for logic relevant to the verification (originating from DNS zone configurations), rather than blindly collecting all code paths ahead of time. Iceberg has been applied to 4 open-source DNS engine implementations (CoreDNS, Bind 9, PowerDNS, and HickoryDNS), revealing 12 new bugs while keeping the spec-to-code ratio below 1:10.

TEACHING

- **TA, Operating Systems (Honor Track)** 2025 Spring
- **TA, Introduction to Computer System** 2023 Fall
- **TA, Introduction to Computer System** 2022 Fall

SERVICES

- **Artifacts Evaluation Committee** SOSP 2025

AWARDS & HONORS

- **Excellent Graduate (Peking University)** 2024
- **Merit Student Pacesetter (Peking University)** 2023
- **Huatai Science and Technology Scholarship (Peking University)** 2023
- **Merit Student Pacesetter (Peking University)** 2022
- **Huatai Science and Technology Scholarship (Peking University)** 2022
- **Merit Student (Peking University)** 2021
- **Shenzhen Finance Institute Scholarship (Peking University)** 2021