Shaoting **Feng**

shaoting.feng.2002@gmail.com | shaoting-feng.github.io | Shaoting-Feng | Shaoting Feng

Research Interests

Systems for ML, Computer Networking

Education

University of Chicago — *Chicago, IL, USA*
Pre-Doctoral MS in Computer Science, advised by Prof. **Junchen Jiang** — *Sep. 2024 - Dec. 2025*

EPFL (Ecole Polytechnique Fédérale de Lausanne) — *Lausanne, Switzerland*
Semester Exchange in Computer Science — *Sep. 2023 - Dec. 2023*

Shanghai Jiao Tong University — *Shanghai, China*
B.E. in Information Engineering, GPA: 3.88, **Ranking: Top 5%**, TOEFL: 108 (R: 30, L: 28, S: 25, W:25) — *Sep. 2020 - Jun. 2024*

Publications

Yuhan Liu, Yuyang Huang, Jiayi Yao, **Shaoting Feng**, Zhuohan Gu, Kuntai Du, Hanchen Li, Yihua Cheng, Junchen Jiang, Shan Lu, Madan Musuvathi, Esha Choukse. *DroidSpeak: KV Cache Sharing for Cross-LLM Communication and Multi-LLM Serving.* [pdf] *NSDI'26*

Siddhant Ray, Rui Pan, Zhuohan Gu, Kuntai Du, **Shaoting Feng**, Ganesh Ananthanarayanan, Ravi Netravali, Junchen Jiang. *METIS: Fast Quality-Aware RAG Systems with Configuration Adaptation.* [pdf] [poster] *SOSP'25*

Shaoting Feng, Yuhan Liu, Hanchen Li, Xiaokun Chen, Samuel Shen, Kuntai Du, Zhuohan Gu, Rui Zhang, Yuyang Huang, Yihua Cheng, Jiayi Yao, Qizheng Zhang, Ganesh Ananthanarayanan, Junchen Jiang. *EVICPRESS: Joint KV-Cache Compression and Eviction for Efficient LLM Serving.* *Submitted to OSDI'26 in an anonymous name*

Shaoting Feng, Hanchen Li, Kuntai Du, Zhuohan Gu, Yuhan Liu, Jiayi Yao, Siddhant Ray, Samuel Shen, Yihua Cheng, Ganesh Ananthanarayanan, Junchen Jiang. *AdaptCache: KV Cache Native Storage Hierarchy for Low-Delay and High-Quality Language Model Serving.* [pdf] [slides] *SOSP workshop BigMem'25*

Yihua Cheng*, Yuhan Liu*, Jiayi Yao*, Yuwei An, Xiaokun Chen, **Shaoting Feng**, Yuyang Huang, Samuel Shen, Kuntai Du, Junchen Jiang. *LMCache: An Efficient KV Cache Layer for Enterprise-Scale LLM Inference.* [pdf] *ArXiv*

Shaoting Feng, Quanquan Peng, Qinya Li, Fan Wu, Guihai Chen. *Grouping Algorithms for Optimal Configuration of Virtual Links in AFDX.* [pdf] [codes] [slides] *Journal of Computer Science and Technology 2025*

Shaoting Feng, Qinya Li, Yaodong Yang, Fan Wu, Guihai Chen. *GIPUT: Maximizing Photo Coverage Efficiency for UAV Trajectory.* [pdf] [codes] *APWeb-WAIM'24*

Open Source Contributions

LMCache (Stars 6.5k): the best performance KV caching layer between LLM inference engines and storage backends. *UChicago*

- Contributed 74 commits (+11,164 / -3,251 LOC), ranking **5th** in total contributions.
- Developed prefill decode disaggregation to reduce tail latency, achieving 20× faster KV cache transmission over vLLM.
- Developed zero-copy CPU offloading for jointly managing GPU and CPU memory, achieving 2.29× TTFT improvement over vLLM.
- Developed multimodal KV cache offloading to accelerate image, video, and audio inference, achieving 5.49× TTFT improvement.
- **Impact:** Widely used in enterprise settings (e.g., NVIDIA, IBM Cloud). >300TB KV cache data + 1.28 billion hit tokens weekly.

vLLM production stack (Stars 2.1k): cluster-wide Kubernetes orchestration. 55 commits (+4,262 / -1,044 LOC), ranking **3rd**. *UChicago*

LMBenchmark: systematic and comprehensive benchmarks for LLM systems. *UChicago*

Selected Projects

EVICPRESS: the first system that jointly considers both lossy compression and eviction for a multi-tier KV cache system. *UChicago*

- Advised by Prof. **Junchen Jiang** and **Ganesh Ananthanarayanan**.
- Proposed a utility function that quantifies the effect of lossy compression and eviction on both quality and TTFT to make decisions.
- Reduced TTFT by 1.43–3.77× with the same quality and improved throughput by 2.0–3.6× at a quality score target of 80%.
- Submitted to **OSDI'26**, accepted by **SOSP workshop BigMem'25**, and implemented in LMCache.

Packet-level fairness metric: implemented in ns-3 and validated in dynamic data center networks. *University of Pennsylvania*

- Advised by Prof. **Vincent Liu** and Dr. **Liangcheng Yu**.

Awards

2024	MPCS Merit-Based Scholarship (Issued by UChicago Pre-Doctoral MS Program)
2023	Dennis C.C. Chan Scholarship (Awarded to 6 undergraduates out of 14K undergraduates university-wide)
2022	Shanghai Government Scholarship (Awarded to **top 0.175%** undergraduates across Shanghai)
2021-2023	Shanghai Jiao Tong University Outstanding Scholarship (Awarded to top 10% undergraduates university-wide)
2019	Silver Medal at Chinese Mathematical Olympiad
2018, 2019	Silver Medal at Chinese Physics Olympiad
2018	Global Top 100 Teams, Regional Top 10 Teams in the American Regions Mathematics League Local mathematics competition

Presentations

Run Multi-Modality Models with LMCache

- SIGCOMM 2025 Full-day Tutorial: Networking for Stateful LLM Inference [talk video] [slides], Sep. 2025

Skills

Programming Python, C/C++, Matlab, VHDL, Verilog

Tools vLLM, kubernetes, FPGA, Network Simulator 3 (NS-3)