KEXIN CHU

11-J Forest Road, Talcott, Farmington, Connecticut 06032, USA
Tel: (+1) 959-995-3930; Email:kexin.chu@uconn.edu
Homepage: https://kexinchu.github.io/

PROFILE

Ph.D. student and experienced software engineer with a strong background in software architecture and distributed systems, proficient in Python (and C++). Over 2 years of combined research and development experience focusing on machine learning systems – particularly Large Language Models (LLMs) – including work on model inference optimization, Mixture-of-Experts (MoE) acceleration, and multi-tier KV-cache management. Adept at using frameworks like PyTorch, vLLM, and SGLang to develop and serve LLMs. Four years of industry experience at Baidu building large-scale services have honed my engineering efficiency and teamwork skills.

EDUCATION

University of Connecticut, US CT *Ph.D. in Computer Science & Engineering (August 2024 - Present)*
Research focus: Efficient large-scale machine learning systems (LLM serving optimizations).
Hefei University of Technology, China *M.S. in Electrical Engineering (August 2017 - June 2020)*
Thesis focus: Computer architecture and AI acceleration.
Hefei University of Technology, China *B.S. in Electrical Engineering (August 2013 - June 2017)*
Specialization: Digital circuit design and embedded systems.

RESEARCH EXPERIENCE

SafeKV & System-Level Co-Design of Privacy Enforcement and KV-Cache Management *(Apr 2025 – Present)*

Developing a unified framework to secure and optimize KV-cache sharing in multi-tenant LLM serving. A concise paper on our core design has been accepted at *MLArchSys'25|ISCA'25*. The full paper is Under review by MLSys 2026.

FlexMoE: Adaptive Expert Prefetching and Cache-Aware Routing for Fast MoE Inference *Oct 2024 - Jun 2025*

Developed a dynamic proactive caching framework to accelerate inference for MoE-based large language models. This system adaptively adjusts expert prefetching based on runtime statistics (e.g. transfer bandwidth, parameter sizes, token-level signals) using a hybrid cross-layer prediction mechanism. Results: Reduced waiting latency to <2% of the baseline and improved expert prediction accuracy by 30%+ across multiple MoE models. Under review at the *(MLSys 2026)*.

Reusing KV-Cache Across Multiple Requests for LLM *Nov 2023 - Aug 2024*

Investigated redundancy in the prefill phase of LLM inference caused by repeated prefix token computation across requests. Proposed **MCaM**, a multi-tier KV-cache management strategy that eliminates redundant computation by reusing cached key–value pairs for identical prefixes and offloads cache storage to host memory to save GPU space. Results: Achieved a 60% reduction in prefill latency. Accepted to the *45th IEEE International Conference on Distributed Computing Systems Conference(ICDCS'25)*.

WORK EXPERIENCE

Baidu, Inc; Senior Software Engineer; Beijing, China July 2020 - August 2024

- Lead developer for Search Push and Recommendation Service (Oct 2023 - Aug 2024).
- Lead developer for Search DeepQA Service (Jul 2020 - Oct 2023).
- Promoted in October 2021 (T3->T4) and July 2023 (T4->T5) because of my exceptional performance (consistent top 30% (M+) in 2021 and 2022).

- **Lead Developer, Search Push & Recommendation Service (2023–2024)**:
  Built a large-scale content recommendation pipeline from the ground up. Optimized data flow (cutting end-to-end latency from 24 hours to 30 minutes) and restructured system architecture for scalability, driving $\tilde{6}$00 million daily data distributions.
- **Lead Developer, Search DeepQA Service (2020–2023)**:
  Overhauled Baidu Search's question-answering system to improve performance and incorporate AI features. Reconstructed the retrieval framework and migrated key components to Golang for higher throughput, supporting 150+ million daily user queries with significantly improved reliability.
- **LLM Integration Project – Wenxin Yiyan (ERNIE Bot) Access System (2023)**:
  Developed and maintained the user access management platform for Baidu's large language model integration into Search. Implemented permission controls, invitation workflows, and activation processes to safely deploy the LLM to users—supporting over 2 million users since launch and enabling new AI-driven search capabilities.
- **Streaming Data Indexing System (2021–2022)**:
  Engineered a hybrid streaming/batch indexing system to accelerate search index updates. Reduced indexing time from 48 hours to 5 minutes and shortened validation cycles from 2 weeks to 1 day, greatly improving data freshness in search results.

PUBLICATION

Kexin Chu, et.al. MCaM: Efficient LLM Inference with Multi-tier KV Cache Management (ICDCS '25)

Kexin Chu, et.al. FlexMoE: Adaptive Expert Prefetching and Cache-Aware Routing for Fast MoE Inference (MLSys '26 under review)

Kexin Chu, et.al. SafeKV: System-Level Co-Design of Privacy Enforcement and KV-Cache Management for LLM Serving (MLsys '26 under review)

Kexin Chu, et.al. M-CANS: Multi-Tier Co-Designed ANNS System for Scalable Multi-Modal Retrieval (AOMC '25 | ISCA '25)

TECHNICAL SKILLS

Programming	Python(*advanced*), Golang, C++
Machine Learning	LLM, MoE, LoRA, et.al
Tools & Frameworks	vLLM, sglang, transformers, PyTorch, Linux, Git

AWARDS

Predoctoral Fellowship, University of Connecticut (2025)

Baidu Pride Special Award, Baidu.Inc (2022)

Breakthrough Innovation Award, Baidu.Inc (2021-2022)

National Scholarship Award, Heifei University of Technology (2018, 2019)