Kan Wu

Operating Systems • Machine Learning Systems
https://cs.wisc.edu/~kanwu • kanwu.wisc@gmail.com • +1 (608) 960-6750 • San Francisco, CA

PROFESSIONAL EXPERIENCE

xAI — Member of Technical Staff MAY 2025 – PRESENT

- Lead applied inference team; overview large-scale inference infrastructure (load balancing, long context-serving, service-level specialization, auto-scaling).
- Develop and optimize SGLang inference engine for reliability, tail performance, observability, and priority-based scheduling/caching.
- Speculative decoding (training and inference) and constrained decoding for production and RL.
- Contribute to major Grok model releases and products, including Grokipedia, X Search, and batch API.

SystemsResearch@Google — Senior Software Engineer SEP 2022 – MAY 2025

- Gemma3n sparse attention (published in NeurIPS 2025)
- Gemini 2.5 pro long-context serving, with long-short context segregation and specialization.
- Conducted research on Google data center memory efficiency, including eBPF based fleet-wide memory-tiering systems (USENIX ATC 2025) and fleet-wide hugepage efficiency (techniques upstreamed to LLVM).

EDUCATION

2016-2022 **UNIVERSITY OF WISCONSIN – MADISON** MADISON, WI
Ph.D. in Computer Science. Advisors: Andrea Arpaci-Dusseau, Remzi Arpaci-Dusseau
Areas: Caching Systems, Storage Systems, Databases

2012-2016 **UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA** HEFEI, ANHUI
B.Sc. in Computer Science (with honor)

SELECTED PUBLICATIONS

[1] **Spark Transformer: Reactivating Sparsity in FFN and Attention**.
Chong You*, **Kan Wu***(co-first author), Zhipeng Jia*, Lin Chen*, Srinadh Bhojanapalli, Jiaxian Guo, Utku Evci, Jan Wassenberg, Praneeth Netrapalli, Jeremiah J. Willcock, Suvinay Subramanian, Felix Chern, Alek Andreev, Shreya Pathak, Felix Yu, Prateek Jain, David E. Culler, Henry M. Levy, Sanjiv Kumar.
NeurIPS'2025, techniques shipped in Google Gemma-3n models

[2] **PageFlex: Flexible and Efficient User-space Delegation of Linux Paging Policies with eBPF.**
Anil Yelam, **Kan Wu*(**corresponding author), Zhiyuan Guo, Rajath Shashidhara, Stanko Novakovic, Suli Yang, Wei Xu, Alex C. Snoeren, Kimberly Keeton.
USENIX ATC'2025, techniques landed in Google fleet memory swapping systems

[3] **The Storage Hierarchy is Not a Hierarchy: Optimizing Caching on Modern Storage Devices with Orthus.**
Kan Wu, Zhihan Guo, Guanzhou Hu, Kaiwei Tu, Ramnatthan Alagappan, Rathijit Sen, Kwanghyun Park, Andrea Arpaci-Dusseau, Remzi Arpaci-Dusseau.
USENIX FAST'2021

[4] **Arya: Arbitrary Graph Pattern Mining with Decomposition-based Sampling.**
Zeying Zhu*, **Kan Wu*** (co-first author), Zaoxing Liu.
USENIX NSDI'2023

[5] **NyxCache: Flexible and Efficient Multi-tenant Persistent Memory Caching.**
Kan Wu, Kaiwei Tu, Yuvraj Patel, Rathijit Sen, Kwanghyun Park, Andrea Arpaci-Dusseau, Remzi Arpaci-Dusseau.
USENIX FAST'2022

PROFESSIONAL SERVICES

Program committee: OSDI'26, MLSYS'26, FAST'25, ATC'25, ATC'24, HotStorage'24

Proceedings chair: SOSP'24

OTHER PUBLICATIONS

[6] **Getting the MOST out of your Storage Hierarchy with Mirror-Optimized Storage Tiering.**
Kaiwei Tu, **Kan Wu**, Andrea Arpaci-Dusseau, Remzi Arpaci-Dusseau.
USENIX FAST'2026

[7] **FineMem: Breaking the Allocation Overhead vs. Memory Waste Dilemma in Fine-Grained Disaggregated Memory Management.**
Xiaoyang Wang, Yongkun Li, **Kan Wu**, Wenzhe Zhu, Yuqi Li, Yinlong Xu.
USENIX OSDI'2025

[8] **SLAP: Segmented Reuse-Time-Label Based Admission Policy for Content Delivery Network Caching**.
Ke Liu, **Kan Wu**, Hua Wang, Ke Zhou, Peng Wang, Ji Zhang, Cong Li.
ACM TACO'2024

[9] **SLAP: An Adaptive, Learned Admission Policy for Content Delivery Network Caching.**
Ke Liu, **Kan Wu**, Hua Wang, Ke Zhou, Ji Zhang, Cong Li.
IPDPS'2023

[10] **WiscSort: External Sorting for Byte Addressable Storage.**
Vinay Banakar, **Kan Wu**, Yuvraj Patel, Kimberly Keeton, Andrea Arpaci-Dusseau, Remzi Arpaci-Dusseau.
VLDB'2023

[11] **Cornus: Atomic Commit for Cloud DBMS with Storage**.
Zhihan Guo, Xinyu Zeng, **Kan Wu,** Wuh-Chwen Hwang, Ziwei Ren, Xiangyao Yu, Mahesh Balakrishnan, Philip A. Bernstein.
VLDB'2022

[12] **Releasing Locks As Early As You Can: Reducing Contention of Hotspots by Violating Two-Phase Locking.**
Zhihan Guo, **Kan Wu**, Cong Yan, Xiangyao Yu.
SIGMOD'2021

[13] **Read as Needed: Building WiSER, a Flash-Optimized Search Engine**.
Jun He, **Kan Wu,** Sudarsun Kannan, Andrea Arpaci-Dusseau, Remzi Arpaci-Dusseau.
USENIX FAST'2020

[14] **Towards an Unwritten Contract of Intel Optane SSD**.
Kan Wu, Andrea Arpaci-Dusseau, Remzi Arpaci-Dusseau.
USENIX HotStorage'2019

[15] **Exploiting Intel Optane SSD for Microsoft SQL Server**.
Kan Wu, Andrea Arpaci-Dusseau, Remzi Arpaci-Dusseau.
ACM DaMoN@SIGMOD'2019