Chihan Cui

chihan.cui@wisc.edu | GitHub | LinkedIn

EDUCATION

University of Wisconsin-Madison — Sep 2025 - 2030 (Expected)
Ph.D. in Computer Science — Madison, WI
Advisor: Ming Liu

University of Toronto (St. George Campus) — Sep 2021 - Mar 2025
B.S. in Computer Science — Ontario, Canada
Advisors: Qizhen Zhang and Phil Bernstein

PUBLICATIONS

CommBench: Can LLMs Write Correct and Efficient GPU Communication Code?
Shuang Ma, Yuyi Li, Yihan Zhang, Hezhi Xie, Danyang Chen, Shuyang Ji, Ziming Mao, Cheng Ji, Ansha Prashanth, Wenting Yang, Yiran Wang, **Chihan Cui**, Pei Yu Lin, Ion Stoica, and Yang Zhou.
arXiv preprint, 2026 | arXiv:2608.04450

Expert-Choice Routing Enables Adaptive Computation in Diffusion Language Models
Shuibai Zhang*, Caspian Zhuang*, **Chihan Cui***, Zhihan Yang, Fred Zhangzhi Peng, Yanxin Zhang, Haoyue Bai, Zack Jia, Yang Zhou, Guanhua Chen, and Ming Liu.
Conference on Language Modeling (**COLM 2026**) | arXiv:2604.01622

UCCL-EP: Portable Expert-Parallel Communication
Ziming Mao, Yihan Zhang, **Chihan Cui**, Kaichao You, Zhongjie Chen, Zhiying Xu, Scott Shenker, Costin Raiciu, Yang Zhou, and Ion Stoica.
USENIX Symposium on Operating Systems Design and Implementation (**OSDI 2026**) | arXiv:2512.19849

Making Sense of DPU Performance for Cloud Data Processing
Jiasheng Hu, **Chihan Cui**, Yuanfan Chen, Philip A. Bernstein, Jialin Li, and Qizhen Zhang.
ACM Symposium on Cloud Computing (**SoCC 2026**) | arXiv:2504.05536

* Equal contribution.

RESEARCH EXPERIENCE

Expert-Choice Routing for Diffusion Language Models
Co-first author | COLM 2026

- Co-first-authored research on expert-choice routing for diffusion language models, enabling balanced expert utilization and adaptive computation.
- The approach varies expert capacity across denoising steps, improving throughput and convergence over token-choice routing.

RESEARCH EXPERIENCE (CONTINUED)

UCCL-EP (1,200+ GitHub stars) Jun 2025 - Dec 2025

Supervisor: Yang Zhou

- Built a portable MoE communication library across NVIDIA/AMD GPUs and heterogeneous NICs by decoupling GPU control from data transmission.
- Integrated with vLLM and Megatron-LM, achieving up to **2.1x** dispatch/combine throughput on AWS EFA and **45%** higher DeepSeek-V3 training throughput on AMD clusters.

dpBento (SoCC 2026) Mar 2024 - Sep 2024

Supervisor: Qizhen Zhang & Philip A. Bernstein

- Built a flexible framework to benchmark SmartNICs and DPUs using database workloads and FIO storage tests.
- Adapted applications across DPU platforms to support comparable performance evaluation in disaggregated environments.

OPEN-SOURCE PROJECTS

Sirius (800+ GitHub stars) Sep 2025 - Present

- Optimized Top-N query execution in a GPU-accelerated SQL engine, identifying global sorting as a major performance bottleneck for interactive analytics.
- Built a hybrid execution engine ("DualWield") that adaptively switches between Heap Sort and Radix Sort based on query limit size and data types.
- Accelerated ClickBench query performance by up to **6.12x** (end-to-end) and **53x** (operator-level) compared with the baseline Sirius engine.

PROFESSIONAL ACTIVITIES

Teaching Assistant, Intro to Big Data Systems Spring 2026
University of Wisconsin-Madison

Teaching Assistant, Intro to Big Data Systems Fall 2025
University of Wisconsin-Madison

TECHNICAL SKILLS

Languages: Python, C/C++, Java.
Tools & Systems: RDMA, Docker, DPDK, PyTorch, CUDA, vLLM, SGLang.