Jianru Ding

John Crerar Library 381, 5730 S Ellis Ave, Chicago, IL 60637 USA

Email: jrding@uchicago.edu   Website: https://jianruding.com

RESEARCH INTERESTS

My research interests lie in the intersection of Computer Architecture and Operating Systems, Machine Learning, and HPC. My work focuses on adapting architectural design and computing resources to large-scale workload fluctuations to accelerate the execution of different workload types. Most recently, my work has focused on adapting LLM-serving frameworks and agentic workloads to hardware resources to provide more computing power in a user-friendly way.

EDUCATION

University of Chicago *Sept 2020 – Present*

Ph.D. and MS in Computer Science. Advisory: Prof. Hank Hoffmann

The Ohio State University *Aug 2016 – Aug 2020*

BS in Computer Science and Finance. Advisory: Prof. Christopher Stewart

- Honor Engineering Program
- Honor Thesis: Characterizing Service Level Objectives
- Graduated with Cum Laude Honor

PUBLICATIONS AND PREPRINTS

[1] Andronicus Rajasukumar, Ruiqi Xu, Tianchi Zhang, Yuqing Wang, Su Tianshuo, Marziyeh Nourian, **Jianru Ding**, Jiya Su, Rajat Khandelwal, Alexander Fell, David F. Gleich, Yanjing Li, Henry Hoffmann, and Andrew A. Chien. Efficient manycore based on many threading and scalable memory parallelism. In *Proceedings of the 40th ACM International Conference on Supercomputing*, ICS '26, Belfast, Northern Ireland, United Kingdom, 2026. Association for Computing Machinery

[2] Ran Zhou, **Jianru Ding**, Chenfeng Gao, Wanli Qian, Benjamin Erickson, Madeline Balaam, Daniel Leithinger, and Ken Nakagaki. Shape-kit: A design toolkit for crafting on-body expressive haptics. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–26, 2025

[3] **Jianru Ding**, Yuqing Wang, Ahsan Pervaiz, Tianchi Zhang, Ruiqi Xu, Jiya Su, Ziyi Zhang, Andrew A. Chien, and Hank Hoffmann. Scalable runtime scheduling for streaming and in-memory tasks, 2025. *In submission*

[4] Andronicus Rajasukumar, Jiya Su, Yuqing, Wang, Tianshuo Su, Marziyeh Nourian, Jose M Monsalve Diaz, Tianchi Zhang, **Jianru Ding**, Wenyi Wang, Ziyi Zhang, Moubarak Jeje, Henry Hoffmann, Yanjing Li, and Andrew A. Chien. Updown: Programmable fine-grained events for scalable performance on irregular applications, 2024. *In submission*. Available at https://arxiv.org/abs/2407.20773

[5] Yuqing Wang, Andronicus Rajasukumar, Tianshuo Su, Marziyeh Nourian, Jose M Monsalve Diaz, Ahsan Pervaiz, **Jianru Ding**, Charles Colley, Wenyi Wang, Yanjing Li, et al. Efficiently exploiting irregular parallelism using keys at scale. In *2023 Workshop on Languages and Compilers for Parallel Computing*, 2023

[6] **Jianru Ding** and Henry Hoffmann. Dps: Adaptive power management for overprovisioned systems. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, SC '23, New York, NY, USA, 2023. Association for Computing Machinery

[7] **Jianru Ding**, Ruiqi Cao, Indrajeet Saravanan, Nathaniel Morris, and Christopher Stewart. Characterizing service level objectives for cloud services: Realities and myths. In *2019 IEEE International Conference on Autonomic Computing (ICAC)*, pages 200–206, 2019

[8] Nathaniel Morris, Indrajeet Saravanan, Pollyanna Cao, **Jianru Ding**, and Christopher Stewart. Slo computational sprinting. In *Proceedings of the ACM Symposium on Cloud Computing*, SoCC '18, page 510, New York, NY, USA, 2018. Association for Computing Machinery

INDUSTRY EXPERIENCE

Research Intern *Seattle, US*
Azure Research - Systems, Microsoft *Jun 2025 – Sep 2025*

- Proposed a hardware-software co-designed hybrid cooling scheme for next-generation AI datacenters
- Profiled LLM inferencing characteristics on latest GPUs and engineered the vLLM serving engine in-depth to support accurate and fine-grained profiling.
- Developed a datacenter-scale cooling simulator that accurately estimates datacenter operating efficiency with dynamic LLM serving traces and climate conditions

Tech Intern *Hefei, China*
iFlytek *May 2019 – Aug 2019*

- Developed the customized deep learning-based speech detection system for clients
- The speech detection system reaches keyword and semantic detection accuracy of more than 99% for clients with expertise in different fields
- The final system was actively adopted by several companies

HIGHLIGHT RESEARCH PROJECTS

Boosting LLM Reasoning with Premise Equilibrium *Jan 2026 – Present*
University of Chicago

- Proposed a new language model training methodology via premise reordering to improve reasoning ability
- Applied parameter-efficient fine-tuning (PEFT) to five open-source LLMs, yielding significant accuracy improvements on reasoning benchmarks

LLM-based Agent serving on heterogeneous GPUs
University of Chicago *Jun 2025 – Present*

- Performed a thorough characterization of agentic LLM workloads regarding power and performance
- Proposed an adaptive agent-serving framework that dynamically disaggregates the prefill and decode phases
- Integrated prefix-caching and LMCache prefill-decode disaggregation in the vLLM engine implementation
- Developed a scalable, Kubernetes-deployable agent-serving framework with adaptive workload placement based on GPU characteristics, improving power efficiency and service-level objective (SLO) satisfaction

The UpDown System Project
University of Chicago *Jun 2022 – Present*

- Leader of a team on developing and validating the UpDown architecture, ISA, network, assembler, compiler, a timing-accurate runtime simulator and a GEM-5 based cycle-accurate runtime simulator
- Proposed the first-of-its-kind million-scale distributed task load-balancer (in submission)
- Designed and implemented a fine-grained vertex-centric distributed GCN training framework (in submission)
- Co-designer of the load balancing scheme for KV Map Shuffle Reduce (ICPC '23)

DPS: Adaptive Power Management for Overprovisioned Systems
University of Chicago *Sept 2020 – Apr 2023*

- Designed the first-of-its-kind model-free stateful power management system for overprovisioned clusters, yielding close performance to optimal model-based approaches and outperforms SLURM by up to 12.4%
- Developed and released the power management program as open source

Cache-based Computational Sprinting respecting industry-level SLOs
The Ohio State University *Apr 2018 – Jun 2020*

- Developed a CNN-based SLO computational sprinting model that increased concurrent cache usage and throughput, reducing SLO–latency slack from 20% to 1%
- Designed a well-defined repeatable Systematic Literature Review (SLR) process for data mining Service Level Objectives (SLO) that reduces potential bias within large-amount literature reviews
- Applied the SLR to accumulate more than 80 sets of Service Level Objective (SLO) samples by datamining more than 50 industrial products and 9,500 published articles

INDEPENDENT PROJECTS

Occupant Well-being Project

- Developed an Electrocardiography analyzing system using deep learning to detect driver emotions in a Honda R&D sponsored project

Decoupled Neural Interfaces for Residual Neural Network using Synthetic Gradients

- Developed an RNN-based decoupled neural interface that takes residues into its prorogation prediction model reducing GPU memory usage by roughly 50

X86 Processor

- Implemented a 32-bit 5-staged x86 processer in C as part of a course project

TEACHING

Teaching Assistant, University of Chicago, Chicago, IL

Parallel Computing *Spring 2021*
Computer Architecture for Scientists *Winter 2020*
Computer Architecture *Fall 2020*

Teaching Assistant, The Ohio State University, Columbus, OH

Intro to Database System *Fall 2019*
Principle of Programming Languages *Fall 2018*
Systems I: Introduction to Low-Level Programming and Computer Organization *Spring 2018*

PROFESSIONAL SERVICE

Reviewer for Transactions on Sustainable Computing 2025

TECHNICAL SKILLS

Skills: LLM Post-Training, Low-level development, Parallel computing, Data mining, Machine learning

Languages: Python, C, C++, Java, SQL

Frameworks & APIs: RestAPI, vLLM, MapReduce, Apache Spark, PyTorch, Kubernetes