Jianru (Jerry) Ding

John Crerar Library 381, 5730 S Ellis Ave, Chicago, IL 60615 USA

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

RESEARCH INTERESTS

My research interest lies in the intersection of HPC, Computer Architecture and Operating Systems, and Machine Learning. My work focuses on adapting computing resources to large-scale workload fluctuations to sprint LLM training and serving. More specifically, I'm interested in how to adapt training and inference pipeline scheduling to a heterogeneous distributed environment to meet higher-level user-defined goals.

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, 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

[2] Yuqing Wang, Andronicus Rajasukumar, Tianshuo Su, Marziyeh Nourian, Jose M Monsalve Diaz, Ahsan Pervaiz, **Ding, Jianru**, 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

[3] **Ding, Jianru** 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

[4] **Ding, Jianru**, 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

[5] Nathaniel Morris, Indrajeet Saravanan, Pollyanna Cao, **Ding, Jianru**, 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

RESEARCH EXPERIENCE

The UpDown System Project *June 2022 – Present*

University of Chicago

- Proposed the first-of-its-kind million-scale distributed task load-balancer (ongoing)
- Designed and implemented a fine-grained vertex-centric distributed GCN training framework (ongoing)
- 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
- 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, which yields 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
The Ohio State University *Apr 2018 – June 2020*

- Developed a CNN and gcForest based Service Level Objective (SLO) computational sprinting modeling approach increasing concurrent cache usage and throughput
- The approach reduced slack between SLOs and application latency from 20% to 1%

Characterizing Service Level Objectives
The Ohio State University *Aug 2018 – Jan 2019*

- 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

INDUSTRY EXPERIENCE

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

INDIPENDENT PROJECTS

Occupant Wellbeing 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*

TECHNICAL SKILLS

Skills: Kernel & Low-level development, Parallel computing, Data mining, Machine learning

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

Frameworks & APIs: MapReduce, Apache Spark, Tensorflow