Bodun Hu

CONTACT INFORMATION

E-mail: bodunhu@utexas.edu
Website: https://www.bodunhu.com

2317 Speedway
The University of Texas at Austin
Austin, TX 78712 USA

RESEARCH INTERESTS

Systems for ML, Operating System, heterogeneity, ML SW-HW Co-design, Distributed System

EDUCATION

The University of Texas at Austin

Ph.D. in Computer Science
Advisor: Aditya Akella

The University of Texas at Austin

M.S. in Computer Science, May 2021
Advisor: Christopher J. Rossbach

The University of Texas at Austin

B.S. in Computer Science, May 2020 (Research Distinction)

PUBLICATIONS

Bodun Hu, Jiamin Li, Le Xu, Myungjin Lee, Akshay Jajoo, Geon-Woo Kim, Hong Xu, Aditya Akella. 2024. BlockLLM: Multi-tenant Finer-grained Serving for Large Language Models. *Preprint.*

Ajay Jaiswal, **Bodun Hu**, Lu Yin, Yeonju Ro, Shiwei Liu, Tianlong Chen, Aditya Aeklla. 2024. FFN-SkipLLM: A Hidden Gem for Autoregressive Decoding with Adaptive Feed Forward Skipping. *EMNLP 24.*

Bodun Hu, Le Xu, Jeongyoon Moon, Neeraja J. Yadwadkar, Aditya Akella. 2024. MOSEL: Inference Serving Using Dynamic Modality Selection. *EMNLP 24.*

Henrique Fingler, Isha Tarte, Hangchen Yu, Ariel Szekely, **Bodun Hu**, Aditya Akella, Christopher J. Rossbach. Towards a Machine Learning-Assisted Kernel with LAKE. *Proceedings of the International Conference on Architectural Support for Programming Languages and Operating System (ASPLOS).*

Bodun Hu and Christopher J. Rossbach. 2020. Altis: Modernizing GPGPU Benchmarks. *Proceedings of the IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS).*

RESEARCH EXPERIENCE

The University of Texas at Austin (UT Austin), Austin, TX, USA.

Research Assistant **2021 - Current**
Implemented efficient mutlimodal model inference system using learning-based optimization technique.
Designed dynamic memory management techniques to optimize the performance of sparse Llama-2 model.

Intel, San Jose, CA, USA.

P4 Dataplane Intern **2022**
TCP-INT: Improved Network Telemetry in TCP Transport for better e2e visibility and improved closed-loop control of TCP workloads.

The University of Texas at Austin (UT Austin), Austin, TX, USA.

Research Assistant **2017 - 2021**

LAKE: Built a generic API remoting system to expose accelerator APIs to OS kernel with close-to-native performances.
ALTIS: Designed a benchmark with improved diversity over existing GPU benchmarks by extending application domains with modern CUDA features.

The University of Texas at Austin (UT Austin), Austin, TX, USA.

Rearch Assistant **2020**

TAS: Ported TAS into P4 to facilitate TCP fast-path migration to programmable NICs.

The University of Texas at Austin (UT Austin), Austin, TX, USA.

Rearch Assistant **2016 - 2017**

G-Code-gen: Designed an automated detection system utilizing readily available hardware, which detects and terminates 3D printing processes upon identification of object defects.

Industry Experience

H3C, Chengdu, China.

Software Engineering Intern **2018**

Devised and implemented a highly effective caching strategy, resulting in a significant reduction of video streaming processing latency on Kubernetes cluster by a factor of 3x.

Wisesoft, Chengdu, China.

Software Engineering Intern **2017**

Developed a data preprocessing pipeline for improved audio classification in an air traffic control system.

Honors and Awards

ISPASS Student Travel Award, 2020

Research Distinction by the College of Natural Sciences (UT Austin), 2020.

Teaching

CS378: System For Machine Learning and Big Data (undergrad)
Teaching Assistant, UT Austin, Fall 2024

CS378: Multicore Operating System Implementation (undergraduate)
Teaching Assistant, UT Austin, Spring 2020

Talks

- *Altis: Modernizing GPGPU Benchmarking*, ISPASS'20 (August 2020)
- *Accelerating Kernel Access to Hardware Acceleration*, Texas Systems Symposium (November 2020)

Service

- Junior Graduate Admissions Committee, UT Austin (Janurary 2021)