Baolin Li

Ph.D. Candidate

Boston
Massachusetts
☎ *512-423-9969*
✉ *li.baol@northeastern.edu*
🖱 *https://baolin-li.netlify.app/*

Education

2024 **Northeastern University**, *Boston, MA.*
Ph.D. in Computer Engineering, *GPA 4.0/4.0*
Advisor: Prof. Devesh Tiwari

Focus Systems for Machine Learning; High Performance Computing; Cloud Computing.

Courses Computer Architecture; Algorithms; Data Mining; Combinatorial Optimization; Deep Learning.

2017 **The University of Texas at Austin**, *Austin, TX.*
M.S. in Electrical and Computer Engineering, *GPA 3.8/4.0*
Graduate Teaching Assistant: Introduction to Automatic Control

2015 **The University of Manchester**, *Manchester, UK.*
B.Eng. (honours) in Electrical and Electronic Engineering, *GPA 4.0/4.0 (First Class Honours)*

Work Experience

Summer 2023 **ML System Research Intern**, *Netflix (Los Gatos, CA).*

- Design and implementation of distributed Graph Neural Network (GNN) training framework.
- Addressed the GPU memory bottleneck challenges on large-scale heterogeneous graphs.
- Framework helped bring up a foundational GNN model into production for recommendation.

Summer 2022 **Research Intern**, *Bosch Research (Sunnyvale, CA).*

- Serverless ML inference engine for online image segmentation service using AWS Lambda.
- Design of RL-based cloud-edge collaboration system for real-time object detection.

2017–2019 **System Engineer**, *Silicon Labs (Austin, TX).*

- System verification and validation of ARM-based mixed-signal SoCs for IoT applications.

Awards

2024 Lux. Veritas. Virtus. Exceptional Graduate Student Award, Northeastern University.

2024 Outstanding Graduate Research Award, Northeastern University College of Engineering.

2023 Best Paper Award Winner at ACM HPDC '23.

2022–2024 ACM and IEEE Student Scholarship for SoCC '22 and HPDC '23, IPDPS'24

2020–2022 Best Paper Award Finalists at SC'20, HPEC'21, HPEC'22, and DATE '22.

Invited Talks

2023 **Improving ML System GPU Utilization in a Multi-Tenant Production Environment**
Netflix ML Training Platform, Aug 2023;
Huawei Cloud Research Seminar, Jan 2023

2022–2023 **Leveraging Heterogeneous Hardware Resources for Efficient ML Inference**
UNC Charlotte Data Intelligence Research Seminar, Sep 2023;
MIT Computational Research in Boston&Beyond (CRIBB) Seminar, Aug 2022

Selected Publications

The selected publications are from highly esteemed HPC/Cloud conferences with acceptance rates typically below 25%. See *Google Scholar* for a full list of published articles.

SC 2023 **Clover: Toward Sustainable AI with Carbon-Aware Machine Learning Inference Service**,
Baolin Li, Siddharth Samsi, Vijay Gadepally, Devesh Tiwari.
Proceedings of the 2023 ACM/IEEE International Conference on High Performance Computing, Networking, Storage and Analysis (SC).

SC 2023 **Toward Sustainable HPC: Carbon Footprint Estimation and Environmental Implications of HPC Systems**,
Baolin Li, Rohan Basu Roy, Daniel Wang, Siddharth Samsi, Vijay Gadepally, Devesh Tiwari.
Proceedings of the 2023 ACM/IEEE International Conference on High Performance Computing, Networking, Storage and Analysis (SC).

HPDC 2023 **KAIROS: Building Cost-Efficient Machine Learning Inference Systems with Heterogeneous Cloud Resources**,
Baolin Li, Siddharth Samsi, Vijay Gadepally, Devesh Tiwari.
Proceedings of the 2023 ACM International Symposium on High-Performance Parallel and Distributed Computing (HPDC).
Best Paper Award Winner

SoCC 2022 **MISO: Exploiting Multi-Instance GPU Capability on Multi-Tenant GPU Clusters**,
Baolin Li, Tirthak Patel, Siddharth Samsi, Vijay Gadepally, Devesh Tiwari.
Proceedings of the 2022 ACM Symposium on Cloud Computing (SoCC).

HPCA 2022 **AI-Enabling Workloads on Large-Scale GPU-Accelerated System: Characterization, Opportunities, and Implications**,
Baolin Li, Rohin Arora, Siddharth Samsi, Tirthak Patel, Rohan Basu Roy, Vijay Gadepally, Devesh Tiwari et al.
Proceedings of the 2022 IEEE International Symposium on High Performance Computer Architecture (HPCA).

NAACL 2022 **Great Power, Great Responsibility: Recommendations for Reducing Energy for Training Language Models**,
Joseph McDonald, **Baolin Li**, Nathan Frey, Devesh Tiwari, Vijay Gadepally, Siddharth Samsi.
Proceedings of the 2022 Findings of the North American Chapter of the Association for Computational Linguistics (NAACL).

SC 2021 **Ribbon: Cost-Effective and QoS-Aware Deep Learning Model Inference using a Diverse Pool of Cloud Computing Instances**,
Baolin Li, Rohan Basu Roy, Tirthak Patel, Vijay Gadepally, Karen Gettings, Devesh Tiwari.
Proceedings of the 2021 ACM/IEEE International Conference on High Performance Computing, Networking, Storage and Analysis (SC).

Technical Skills

Programming Python, CUDA, C, C++, C#, MATLAB, Verilog.

Tools and Frameworks Pytorch, Tensorflow, Docker, Jmeter, Flask, Pytorch Gemoetric, DeepSpeed, HuggingFace, gRPC, Ray, Spark, Kubernetes, MPI, OpenMP, Triton, TensorRT, NVIDIA Multi-Process Service, Multi-Instance GPU, Pandas, Scikit-learn, Matplotlib