# 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

Interests High Performance Computing (HPC); Cloud Computing; Machine Learning for Systems and Systems for Machine Learning.

Focus Cost-Effective ML Inference; GPU Multi-Tenancy; Carbon-Aware ML Systems.

Courses Computer Architecture; Algorithms; Data Mining; Combinatorial Optimization; GPU Programming.

**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*

## Work Experience

Summer 2023 **ML System Research Intern, Netflix**, *Los Gatos, CA.*
Responsibility: Machine learning platform for scalable foundation model training.
- Developed a Graph Neural Network (GNN) training framework based on Pytorch Geometric for Netflix's content knowledge graph training.
- Addressed the GPU memory bottleneck and multi-GPU distributed training challenges on large-scale heterogeneous graphs.
- Proved the proposed solution to maintain high embedding quality for downstream tasks while achieving significant speedup over the previous solution (>50x).
- Released GNN training software package for machine learning algorithm researchers.
- Patent filing and technical paper writing.

Summer 2022 **Research Intern, Bosch Research**, *Sunnyvale, CA.*
Responsibility: Cloud system design and implementation for ML applications.
- Deployed a serverless ML inference engine for image segmentation service.
- Designed RL-based edge-cloud inference system for real-time object detection, filed for patent.
- Applied cloud services including AWS EC2, Elastic Container Service, Lambda function, API Gateway, Application Load Balancer, S3 storage, distributed load testing.

2017–2019 **System Engineer, Silicon Labs**, *Austin, TX.*
Responsibility: R&D of ARM-based mixed-signal SoCs for IoT applications.
- Automated testing software development to debug and characterize analog IP blocks.
- SoC chip top verification using digital behavioral models in SystemVerilog.
- System-level benchmarking in different energy modes, product datasheet generation.
- Close cooperation with worldwide teams in system, application, firmware and marketing.

## Awards

- **Lux. Veritas. Virtus. Exceptional Graduate Student Award**
  Northeastern University, 2024.
- **Outstanding Graduate Research Award**
  Northeastern University College of Engineering, 2024.
- **Best Paper Award Winner**
  ACM International Symposium on High-Performance Parallel and Distributed Computing (HPDC 2023).
- **ACM and IEEE Student Travel Scholarships**
  IEEE International Parallel and Distributed Processing Symposium (IPDPS 2024).
  ACM International Symposium on High-Performance Parallel and Distributed Computing (HPDC 2023).
  ACM Symposium of Cloud Computing (SoCC 2022).
- **University Nominee of Google PhD Fellowship**
  Northeastern University, 2022.
- **Best Paper Award Finalist**
  Design, Automation and Test in Europe Conference (DATE 2022).
- **Outstanding Paper Award**
  IEEE High Performance Extreme Computing Conference (HPEC 2022).
  IEEE High Performance Extreme Computing Conference (HPEC 2021).
- **Best Paper Finalist and Best Student Paper Finalist**
  ACM/IEEE International Conference on High Performance Computing, Networking, Storage and Analysis (SC 2020).
- **First Class Honour**
  The University of Manchester

## Invited Talks

- **Data Intelligence Research Lab Seminar**
  **University of North Carolina Charlotte, 2023**
  Building Cost-Efficient Machine Learning Inference Systems with Heterogeneous Cloud Resources
- **Machine Learning Training Platform Design and Strategy**
  **Netflix, 2023**
  Enabling Single-GPU Multi-Tenancy for Efficient Operation of Resource-Limited GPU Cluster
- **Huawei Cloud InnoWave Outstanding Research Paper Seminar**
  **Huawei Cloud, 2023**
  How to Improve GPU Utilization for Machine Learning Systems in a Multi-Tenant Production Environment?
- **COMPUTATIONAL RESEARCH in BOSTON and BEYOND**
  **MIT CRIBB Seminar, 2022**
  Leveraging Heterogeneous Hardware Resources for Efficient Machine Learning Inference Service.

## Selected Publications

Most of the selected publications are from highly esteemed HPC/Cloud conferences with acceptance rates typically below 25%, such as ATC, HPCA, HPDC, SC, and SoCC.
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).*

*SoCC 2023* **Sustainable Supercomputing for AI: Experiences from GPU Power-Capping at HPC Scale**,
Dan Zhao, Siddharth Samsi, **Baolin Li**, Devesh Tiwari, Vijay Gadepally.
*Proceedings of the 2023 ACM Symposium on Cloud Computing (SoCC), in press.*

*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).*

*HPEC 2022* **Benchmarking resource usage for efficient distributed deep learning**,
Nathan C Frey, **Baolin Li**, Joseph McDonald, Dan Zhao, Michael Jones, David Bestor, Devesh Tiwari, Vijay Gadepally, Siddharth Samsi.
*Proceedings of the 2021 IEEE Conference on High Performance Extreme Computing (HPEC).*
***Outstanding Paper Award***

*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).*

*HPEC 2021* **Serving Machine Learning Inference Using Heterogeneous Hardware**,
**Baolin Li**, Vijay Gadepally, Siddharth Samsi, Mark Veillette, Devesh Tiwari.
*Proceedings of the 2021 IEEE Conference on High Performance Extreme Computing (HPEC).*
***Outstanding Student Paper Award***

*ATC 2020* **UREQA: Leveraging Operation-Aware Error Rates for Effective Quantum Circuit Mapping on NISQ-Era Quantum Computers**,
Tirthak Patel, **Baolin Li**, Rohan Basu Roy, Devesh Tiwari.
*Proceedings of the 2020 USENIX Annual Technical Conference (USENIX ATC).*

*SC 2020* **Experimental Evaluation of NISQ Quantum Computers: Error Measurement, Characterization, and Implications**,
Tirthak Patel, Abhay Potharaju, **Baolin Li**, Rohan Basu Roy, Devesh Tiwari.
*Proceedings of the 2020 ACM/IEEE International Conference on High Performance Computing, Networking, Storage and Analysis (SC).*
***Best Paper and Best Student Paper Finalist***

## Open-Source Contributions

- **MISO: improving efficiency of multi-tenant GPU clusters**
  https://doi.org/10.5281/zenodo.7853719
- **Systematic study of carbon footprint in modern HPC systems**
  https://doi.org/10.5281/zenodo.10095592
- **Clover: enabling carbon-aware machine learning inference**
  https://doi.org/10.5281/zenodo.10109954
- **Kairos: exploiting heterogeneous cloud resources for machine learning inference**
  https://doi.org/10.5281/zenodo.7888058
- **Characterization and analysis of the MIT SuperCloud datacenter**
  https://doi.org/10.5281/zenodo.6040279
- **Ribbon: cost-effective and QoS-aware deep learning model inference system**
  https://doi.org/10.5281/zenodo.5262865

## Media Features

- **To 'green' AI, scientists are making it less resource-hungry, Science News Explores**
- **New tools to help reduce the energy that AI models devour, MIT News**
- **Taking a magnifying glass to data center operations, MIT News**
- **Study: Dealing with increasing power needs of ML, Dataconomy**

## Services

2024 Reviewer for IEEE Computer Architecture Letters
2024 Reviewer for IEEE Transactions on Cloud Computing
2023 Reviewer for Journal of Parallel and Distributed Computing
2022 Conference review at IEEE/IFIP Dependable Systems and Networks Conference (DSN)