Ruobing Han

Email: hanruobing@gatech.edu Homepage: https://drcut.github.io/

Education

Georgia Institute of Technology (GT), USA 2021.5 – Present
PhD Candidate, Computer Science (CS)
Research Areas: Compiler, Hardware Architecture
Advisor: Prof. Hyesoon Kim

Peking University (PKU), China 2014.9 – 2018.7
Bachelor of Science, Computer Science (CS)

Internship

Google Sunnyvale, CA, USA

- **TPU Compiler Optimization** (2024.5-2024.7)
  - Work in TPU Performance team and TPU Compiler team.
  - Explore optimal solutions for Memory Space Assignment on TPU hardware.
- **ML Debug Toolkit Development** (2023.5-2023.7)
  - Work in the ML Functional Debugging team.
  - Implement compiler static analysis to detect bugs in Tensorflow programs
- **LLVM Compiler Development** (2022.5-2022.8)
  - Work in the LLVM Core team.
  - Implement Loop Unswitching Transformation with LLVM Function Pass.

Research

Migrating CUDA to non-NVIDIA Devices

- Build a framework to execute unmodified CUDA source code on CPUs (x86, ARM, and RISC-V) and AMD GPUs.
- The related papers are published in IEEE/ACM International Symposium on Microarchitecture (MICRO), ACM Transactions on Architecture and Code Optimization (TACO), and ACM Transactions on Design Automation of Electronic Systems (TODAES).
- Project link: `https://github.com/cupbop/CuPBoP`

Improving C++ Incremental Build Execution Time

- Enhance the performance of the incremental build process by recording previous compilation results.
- Develop and implement a proof-of-concept model in the LLVM-14 compiler, achieving a 6.72% speedup on popular C++ projects.
- Published in the Proceedings of The International Symposium on Code Generation and Optimization (CGO) 2024.

Solving the Phase-Ordering Problem with Reinforcement Learning

- Develop a Reinforcement Learning model to address the compiler phase-ordering problem.
- Propose a novel pruning solution that exponentially expands the search space, enabling the Reinforcement Learning model to find optimal solutions in a reasonable time frame.
- The proposed solution generates programs that are 12% faster or 17.6% smaller than the programs produced by LLVM O3/Oz optimizations.
- Published in the Proceedings of The International Conference on Compiler Construction (CC) 2024.

Distributed Training Neural Network with Low-Precision

- Propose an algorithm to avoid overflow while using low-precision floating-point for gradients.
- Use 8-bit floating points to train ResNet50 on a large-scale distributed system.
- Published in the Proceedings of the International Conference on High Performance Computing 2021.

Publications

Conferences

- **Ruobing Han**, Jisheng Zhao, Hyesoon Kim. "Unleashing CPU Potential for Executing GPU Programs through Compiler/Runtime Optimizations" The IEEE/ACM International Symposium on Microarchitecture (MICRO) 2024.
- **Ruobing Han**, Jisheng Zhao, Hyesoon Kim. "Enabling Fine-Grained Incremental Builds by Making Compiler Stateful" The International Symposium on Code Generation and Optimization (CGO) 2024.
- **Ruobing Han**, Hyesoon Kim. "Exponentially Expanding the Phase-Ordering Search Space via Dormant Information" The International Conference on Compiler Construction (CC) 2024.
- **Ruobing Han**, James Demmel, Yang You. "Auto-Precision Scaling for Distributed Deep Learning" International Conference on High Performance Computing 2021.

Journals

- **Ruobing Han**, Jun Chen, Bhanu Garg, Xule Zhou, John Lu, Jeffrey Young, Jaewoong Sim, Hyesoon Kim. "CuPBoP: Making CUDA a Portable Language" ACM Transactions on Design Automation of Electronic Systems (TODAES) 2024.
- **Ruobing Han**, Jaewon Lee, Jaewoong Sim, Hyesoon Kim. "COX: Exposing CUDA Warp-Level Functions to CPUs" ACM Transactions on Architecture and Code Optimization (TACO) 2022.
- Peng Sun, Wansen Feng, **Ruobing Han**, Shengen Yan, Yonggang Wen. "Optimizing Network Performance for Distributed Deep Neural Network Training on GPU Clusters: ImageNet/AlexNet Training in 1.5 Minutes" IEEE Transactions on Big Data 2020.