Yufan Xu

Contact Information

19608 Pruneridge Ave,
Cupertino,
California 95014

Phone: (352) 278-6832
Email: yf.xu@utah.edu
Profile: Google Scholar

Education

University of Utah, Salt lake City, Utah USA (GPA 4.0)

Ph.D, Computer Science, May, 2024

University of Florida, Gainesville, Florida USA (GPA 3.65)

M.S., Computer Science, May, 2016

Soochow University, Suzhou, Jiangsu China (GPA 3.50)

B.Eng., Software Engineering, May, 2014

Publication

Accelerated Auto-Tuning of GPU Kernels for Tensor Computations
ICS 24
- Chendi Li*, Yufan Xu*, Sina Mahdipour Saravani, P. Sadayappan

CoNST: Code Generator for Sparse Tensor Networks
TACO 24
- Saurabh Raje, Yufan Xu, Atanas Rountev, Edward F. Valeev, P. Sadayappan

PEAK: Generating High-Performance Schedules in MLIR
LCPC 23
- Amir Tavakkoli*, Sameeran Joshi*, Shreya Singh, Yufan Xu, P. Sadayappan, Marry Hall

Effective Performance Modeling and Domain-Specific Compiler Optimization of CNNs for GPU
PACT 22
- Yufan Xu, Qiwei Yuan, Erik Curtis Barton, Rui Li, P. Sadayappan, Aravind Sukumaran-Rajam

Training of Deep Learning Pipelines on Memory-Constrained GPUs via Segmented Fused-Tiled Execution
CC 22
- Yufan Xu, Saurabh Raje, Atanas Rountev, Gerald Sabin, Aravind Sukumaran-Rajam, P. Sadayappan

Efficient Distributed Algorithms for Convolutional Neural Networks
SPAA 21
- Rui Li, Yufan Xu, Aravind Sukumaran-Rajam, Atanas Rountev, P. Sadayappan

Analytical characterization and design space exploration for optimization of CNNs
ASPLOS 21
- Rui Li, Yufan Xu, Aravind Sukumaran-Rajam, Atanas Rountev, P. Sadayappan

Dependence-aware, unbounded sound predictive race detection
OOPSLA 19
- Kaan Genç, Jake Roemer, Yufan Xu, Michael D. Bond

Research Experience

Uber Technologies Inc. **May, 2024 - Now**

- Work on memory allocation optimization with PGO in GO compiler
- Work on error propagation fixing with GenAI in GO monorepo
- Work on ML infrastructure inference performance

University of Utah **August, 2019 - May, 2024**

- Worked on search space optimization in TVM
- Worked on design space exploration for optimizing CNN for GPUs
- Worked on memory efficiency for large input on ML system (pytorch)
- Worked on opmin optimization pass for tensor contraction in CCSD benchmark on MLIR
- Worked on a tile-size optimization problem for affine programs in the polyhedral model

The Ohio State University **August, 2017 - May, 2019**

- Worked on data race detection in Java program

Teaching & Advicing

Course Instructor
The Ohio State University **Fall, 2018, Spring, 2019**

- Instructor for two semesters of *CS1223 Introduction to Computer Programming In Java.*
- Taught the general concepts of computer programming and programming languages by providing practical experience programming in the Java.

Teaching Assistant
University of Utah **Spring, 2020**

- Teaching Assistant for *CS 6230 Parallel Computing and HPC.*
- Planned course project, graded assignments and projects.

Working Experience

Uber, Sunnyvale, CA, USA

Software Engineer II May, 2024 - Now

Uber, Sunnyvale, CA, USA

PhD Software Engineer(Intern) May, 2023 - August, 2023

LatentAI, Princeton, NJ, USA

Compiler Engineer(Intern) May, 2022 - August, 2022

T-CETRA, Columbus, OH, USA

Software Engineer(Intern) May, 2019 - August, 2019

Service

Program Committee:
CGO '25
Artifact Evaluation Committee:
ASPLOS '21, '22 ; *CGO* '23, '24 ; *MICRO* '23 ; *CC* '24
Journal reviewer:
ACM Transactions on Architecture and Code Optimization(TACO)
Future Generation Computer Systems
Mentoring:
SIGPLAN-M mentor '24-Now