Chenxi Yang

cxyang@cs.utexas.edu | chenxi-yang.github.io | linkedin.com/in/chenxi-yang-ut | +1 (512) 960-6965

Education

The University of Texas at Austin Aug 2019 – May 2025 (expected)
PhD in Computer Science, advised by Prof. Swarat Chaudhuri Austin, TX

Fudan University Sep 2015 – Jul 2019
BSc in Computer Science (Honors), advised by Prof. Yang Chen Shanghai
- Graduated with Highest Distinction

Work Experience

Google

PhD Intern, System Research Group May 2024 – Aug 2024, Seattle, WA
- **Created TPU Scheduling Simulator**: Led the design and implementation of the first lightweight TPU scheduling simulator within Google. Fully replicated existing TPU scheduling algorithms and automated ML training and inference job scheduling on TPUs, providing a rapid solution to explore TPU design spaces. Tested job distribution ideas for LLM serving systems.
- **Optimize TPUs with Machine Learning (ML)**: Designed and implemented an ML-based, job-lifetime-aware algorithm to optimize TPU utilization by up to 50%, enhancing user experience.
- **Multi-Model Support**: Provided GBT, MLP, and LLM training and serving support for production and internal workloads at Google.

Student Researcher and PhD Intern, Storage Analytics Team May 2023 – Jan 2024, Remote and Sunnyvale, CA
- **Created a New ML-driven Storage Approach**: Led the design and implementation of the first ML-driven, cross-layer storage placement solution in warehouse-scale computers.
- **Workloads Scale & Enhanced Performance**: Enabled the data placement for files from over 500 Google Cloud clusters and provides a 2.48x total cost of ownership savings compared to existing production solutions — estimated to save $12 million upon full deployment.
- **Impact**: First-authored a research paper currently under submission. The solution is being rolled out to production.

Goldman Sachs

Summer Analyst, Engineering Jun 2018 – Aug 2018, Hong Kong
- **Tested High Frequency Trading Systems**: Built a workload generation tool that simulated trading orders flowing through OSI layers to test the new generation ultra-low-latency DMA trading gateway.
- **Impact**: The tool identified critical bugs during the trading system development phase.

Selected Publications

Certified Learning for Congestion Control. [Preprint] Under Review.

A Practical Cross-Layer Approach for ML-Driven Storage Placement in Warehouse-Scale Computers. [Preprint] ***C. Yang****, E. Li, M. Maas, M. Uysal, U. Hafeez, A. Merchant, R. McDougall.*

Safe Neurosymbolic Learning with Differentiable Symbolic Execution [Paper] ***C. Yang****, S. Chaudhuri.* ICLR 2022.

Certifiably Robust Reinforcement Learning through Model-Based Abstract Interpretation. [Paper] ***C. Yang****, G. Anderson, S. Chaudhuri.* SaTML 2024.

On a Foundation Model for Operating Systems. [Paper] *D Saxena, N. Sharma, D. Kim, R. Dwivedula, J.i Chen,* ***C. Yang****, S. Ravula, Z. Hu, A. Akella, J. Biswas, S. Chaudhuri, I. Dillig, A. Dimakis, D. Kim, C. Rossbach.* Neurips 2023, ML for Systems Workshop.

Improved Modeling of RNA-binding Protein Motifs in An Interpretable Neural Model of RNA Splicing. [Paper] *K. Gupta,* ***C. Yang****, K. McCue, O. Bastani, P. A. Sharp, C. Burge, A. Solar-Lezama.* Genome Biology 25 (1) 23. ICML 2023. **Spotlight** (Computational Biology Workshop)

Adaptive Scheduling for Edge-Assisted DNN Serving. [Paper] *J. He,* ***C. Yang****, Z. He, G. Baig, L. Qiu.* MASS 2023.

Accelerating Mobile Applications at the Network Edge with Software-Programmable FPGAs. [Paper] *S. Jiang, D. He,* ***C. Yang****, C. Xu, G. Luo, Y. Chen, Y. Liu, J.Jiang.* INFOCOM 2018.

For a complete list of my publications, please visit my website and my Google Scholar profile.

Selected Projects

TPU Scheduling Optimization May 2024 – Aug 2024
Project lead Google

- **Created TPU Scheduling Simulator**: Led the creation of Google's first lightweight TPU scheduling simulator, replicating existing scheduling algorithms and automating ML training and inference job scheduling on TPUs. Accelerated exploration of TPU design spaces and tested job distribution strategies for LLM serving systems.
- **Enhanced TPU Utilization with ML**: Designed and implemented an ML-based, job-lifetime-aware algorithm, increasing TPU utilization by up to 50% and improving user experience.
- **Supported Multi-Model Training and Serving**: Enabled GBT, MLP, and LLM training and serving support for production and internal workloads.

ML-Driven Storage Placement in Data Center May 2023 – Jan 2024
Project lead Google

- **Pioneered ML-Driven Storage Placement**: Led the design and implementation of the first ML-driven, cross-layer storage placement solution in warehouse-scale computers.
- **Achieved Significant Cost Savings**: Enabled data placement for files from over 500 Google Cloud clusters, resulting in a 2.48x reduction in total cost of ownership compared to existing solutions—estimated to save $12 million upon full deployment.
- **Impact**: First-authored a research paper under submission; the solution is being deployed in production environments.

From Theory to Practice: Certifiably Performant ML systems Jul 2020 – Sep 2024
Project lead UT-Austin

- **Certified Reinforcement Learning for Networked Systems**
  - **Created the System**: Led the design and implementation of the first approach to building ML-controlled systems that integrates learning with formal certification in the loop. This system defines and certifies key system properties related to performance and robustness.
  - **Performance Improvement**: The trained system controllers reduced delay by 78% and yielded better worst-case performance compared to existing methods.
- **DSE: Safe Neurosymbolic Learning with Differentiable Verification**
  - **DSE Algorithm Design**: Developed a pioneering approach for end-to-end, worst-case-safe parameter learning for neural networks within non-differentiable, symbolic programs.
  - **Excellent Safe Performance**: Integrated symbolic execution and stochastic gradient estimators, enabling applications in autonomous driving and critical healthcare.
- **CAROL: Certified Reinforcement Learning**
  - **CAROL Algorithm Design**: Designed the first RL framework with episode-level certifiable adversarial robustness, based on a new combination of model-based learning and abstract interpretation.
  - **Theory and Evaluation**: Provided rigorous theoretical analysis establishing the soundness of CAROL, verified through control benchmarks.

Edge Server DNN Video Processing Acceleration Aug 2019 – Jun 2020
Project contributor UT-Austin

- **Developed Batch-Based Deep Neural Network (DNN) Scheduling Algorithm**: Conceived a batching-aware DNN scheduling methodology to enhance edge DNN request management. Designed and implemented collaborative DNN executions at the edge to accelerate processing on commodity hardware.
- **Performance Improvements**: Reduced completion time and improved system capacity by 20%–400% over baselines when serving DNNs. Increased the on-time ratio, enhancing system capacity by more than 22% over Earliest Deadline First (EDF) with batching strategy.

FPGA-Based Edge Computing for Accelerating Mobile Applications Jul 2017 – Aug 2017
Project contributor Peking University

- **Developed FPGA-Based Edge Computing Model**: Engineered a prototype that minimizes response time and energy consumption for interactive mobile applications by offloading computation to an FPGA-based edge.
- **Performance Improvements**: Achieved up to 3x/15x faster response times over CPU-based edge/cloud offloading and enhanced energy efficiency by up to 29.5%.

Selected Awards

- PLMW@PLDI Scholarship 2022
- Outstanding Undergraduate Graduate, Shanghai Region 2019
- Honors Student Award, Top Talent Undergraduate Program, Fudan University 2019

- National Scholarship, The Ministry of Education of China 2017

Invited Talks & Presentations

- Google Cloud & Google Deepmind, Lifetime-aware TPU Scheduling Aug 2024
- Google SRG Annual Event, Lifetime-aware Bin Packing for TPUs Aug 2024
- SaTML 2024, Certifiably Robust Reinforcement Learning through Model-Based Abstract Interpretation Apr 2024
- Google Deepmind, Learning File Placement Policies in Data Processing Pipelines Aug 2023
- Google Cloud, Learning File Placement Policies in Data Processing Pipelines Aug 2023
- MIT, Safe Neurosymbolic Learning with Differentiable Symbolic Execution Oct 2022
- Caltech, Safe Neurosymbolic Learning with Differentiable Symbolic Execution Jul 2022
- ICLR 2022, Safe Neurosymbolic Learning with Differentiable Symbolic Execution May 2022

Teaching & Service

- Artifact Evaluation Committee ASPLOS 2025 Spring
- Reviewer AIPLANS@Neurips 2021; Neurips 2022, 2023, 2024; ICML 2023, 2024; ICLR 2023, 2024
- Teaching Assistant, CS373: Software Engineering Fall 2019, Spring 2020

Others

- **Programming Languages**: Python, C/C++, Java, Matlab, ACL2 ...
- **Machine Learning**: PyTorch, TensorFlow, PyTorch Lightning, Keras, Scikit-Learn, ...
- **Technical**: Algorithms & Data Structures, Machine Learning Systems (Software & Hardware), Artificial Intelligence, Reinforcement Learning, Formal Verification, ...