QUANG DUONG

CONTACT INFORMATION

Office 5.418D
Gates Dell Complex
Department of Computer Science
The University of Texas at Austin
Austin, Texas 78712

✉ duongquang1@gmail.com
✉ qduong@utexas.edu
🔗 quangduong.me
🐙 quangmire
in duongquang1

RESEARCH INTERESTS

My research aims to rethink the formulation of temporal prefetching to improve the efficiency and performance of both neural and non-neural temporal prefetchers. In addition, I've also delved into applying machine learning techniques to other microarchitectural prediction problems (e.g., cache replacement and branch prediction) to distill insights that can improve the designs of existing predictors.

EDUCATION

The University of Texas at Austin . . . . 2020 - Present
- Ph.D. in Computer Science [GPA: 4.0]
- Advisor: Calvin Lin

The University of Texas at Austin . . . . 2015 - 2020
- B.S./M.S. in Computer Science
- B.S. in Mathematics (Specialization in Scientific Computing)
- *Summa Cum Laude* [GPA: 3.9482]

EXPERIENCE

Graduate Research Assistant . . . . Spring 2021 - Present
Department of Computer Science, The University of Texas at Austin
- Reduced the size ($>1000\times$) and latency ($>100\times$) of neural temporal prefetchers while improving generalization by reformulating the temporal prefetching prediction problem.
- Streamlined on-chip temporal prefetching to improve storage efficiency by 33%.
- Extracted salient branches from profiled neural branch predictors for compression.

CPU Performance Exploration Intern . . . . Summer 2025
Arm, Austin
- Implemented dynamic predictor for modulating front-end speculation, providing a 20:1 energy savings to performance loss trade-off.

CPU Performance Exploration Intern . . . . Summer 2024
Arm, Austin
- Implemented temporal prefetch engine with up to 3% performance headroom.
- Managed prefetcher aggression with fine-grained throttling using machine learning.

CPU Performance Exploration Intern . . . . Summer 2023
Arm, Austin
- Improved speculative cache management using SOTA academic prediction algorithms.
- Explored performance and efficiency headroom of modern cache replacement policies.

Architecture Research Intern . . . . Summer 2022
Arm Research, Austin

- Found double digit traffic reduction opportunity using smarter prefetcher management.
- Implemented practical ML algorithm that achieves a significant portion of the headroom.

Graduate Research Assistant . . . . Fall 2019 - Summer 2020
Space and Geophysics Laboratory, Applied Research Laboratories, Austin

- Proposed and implemented data format for $4\text{-}10\times$ speedup in processing time.
- Migrated code base from Python 2 to 3 and wrapped legacy C++ code into Python libraries with SWIG and Cython to streamline data analysis.

Student Technician . . . . Fall 2018 - Summer 2019
Space and Geophysics Laboratory, Applied Research Laboratories, Austin

- Leveraged clustering and reinforcement learning techniques to iteratively refine estimations of ionospheric model parameters competitive with commercial software.
- Achieved real-time speed by adaptively sampling a fast surrogate model.

Honors Scholar . . . . Summer 2018
Space and Geophysics Laboratory, Applied Research Laboratories, Austin

- Denoised raw ionosonde output via clustering and adaptive thresholding.
- Generated interdependent feasible ranges for the ionospheric model parameters from denoised ionosonde information via multi-layer perceptrons.

Texas Institute for Discovery Education in Science Fellowship . . . . Summer 2017
College of Natural Sciences, The University of Texas at Austin

- Analyzed different Pareto frontiers by weighting the objective function for the 3D model approximation genetic algorithm that traded off generated OpenSCAD code complexity with convergence and error rate to allow for downstream non-expert modification.
- Implemented vectorized voxelization code to approximate expensive objective functions.

Freshman Research Initiative Fellowship . . . . Summer 2016
College of Natural Sciences, The University of Texas at Austin

- Adapted CPPN-NEAT to refine a neural network to functionally approximate a 3D model.
- Generated OpenSCAD's model descriptor language using the neural network output.

PUBLICATIONS ━━

***Streamlined On-Chip Temporal Prefetching* [HPCA]** 2026
- Quang Duong and Calvin Lin

***A New Formulation of Neural Data Prefetching* [ISCA]** 2024
- Quang Duong, Akanksha Jain, and Calvin Lin

TEACHING ━━━

CS 395T: Prediction Mechanisms in Comp Arch [Co-Instructor] . . . . Spring 2026

CS 395T: Prediction Mechanisms in Comp Arch [Co-Instructor] . . . . Spring 2025

CS 395T: Prediction Mechanisms in Comp Arch [Co-Instructor] . . . . Spring 2024

CS 395T: Prediction Mechanisms in Comp Arch [Co-Instructor] . . . . Spring 2023

CS 395T: Prediction Mechanisms in Comp Arch [Co-Instructor] .... Spring 2022

CS 380P: Parallel Systems [TA] .... Summer 2021

CS 380P: Parallel Systems [TA] .... Fall 2020

CS 373: Software Engineering [Undergrad TA] .... Spring 2018

CS 309: Computational Intelligence in Game Research FRI [Mentor] .... Spring 2017

ACTIVITIES

MICRO [Artifact Evaluation Committee Member] 2023

PPoPP [Artifact Evaluation Committee Member] 2022

ML Data Prefetching Competition at MLArchSys / ISCA [Co-Organizer] 2021

ACM [Member] .... Fall 2015 - Present

Machine Learning and Data Science Student Org [Member] .. Spring 2016 - Spring 2018

Information and Systems Security Society [Member] .... Fall 2015 - Spring 2017

AWARDS

Wesley W. Calhoun, Jr. Endowed Scholarship .... Spring 2025 - Spring 2026

UT CS Departmental Student Travel Grant .... Summer 2025

Student Travel Grant .... ISCA 2025

TIDES Advance Summer Research Fellowship .... Summer 2017
  · Awarded to undergraduate students working on advanced research with a supervisor

Freshman Research Initiative Fellowship .... Summer 2016
  · Awarded to freshmen students to continue working on their FRI research proposal

Distinguished College Scholar / Students of High Academic Achievement .. 2017 - 2019

Kemp-Forman Memorial Endowed Presidential Scholarship .... 2017 - 2018

Tracor/Frank McBee, Jr. Scholarship .... 2016 - 2017

RESEARCH TALKS

A New Formulation of Neural Data Prefetching .... Summer 2024
Arm, Austin

SKILLS

  · **Languages**: Python, C/C++, Rust, HTML/CSS, Go, Java, JavaScript, x86 Assembly, LaTeX, SQL, MATLAB, Julia
  · **Libraries**: numpy, PyTorch, TensorFlow, sklearn, scipy, matplotlib/seaborn, React, Flask, OpenGL
  · **Other Skills**: Unix Systems (Ubuntu, Debian, Arch Linux), Vietnamese, 🐱 Cat Enthusiast

MISC

- **Open-Source Contributions**
  - ChampSim
  - GPSTk during my UT ARL graduate assistantship.