Naifeng Zhang

Department of Electrical and Computer Engineering
College of Engineering
Carnegie Mellon University

naifengz@cmu.edu
+1 323 868 5267
naifengz.com

Research Interests: High-performance code generation targeting CPUs, GPUs, and ASICs; semantics lifting for performance portability and safeguarding AI-generated code.

EDUCATION

Ph.D. Electrical and Computer Engineering, Carnegie Mellon University, 2026
Advisor: Franz Franchetti

M.S. Electrical and Computer Engineering, Carnegie Mellon University, 2024
Advisor: Franz Franchetti

B.S. Computer Science, University of Southern California, 2021
Advisor: Viktor K. Prasanna
Thesis: Lightweight Augmented Neural Network For Performance Prediction and Its Applications
W.V.T. Rusch Undergraduate Engineering Honors Program

B.S. Mathematics, University of Southern California, 2021
Departmental Honors Program

RESEARCH APPOINTMENTS

2025 NVIDIA Research
Intern, Programming Systems and Applications Research Group
Santa Clara, United States

AWARDS

2024 Best Poster Runner-up
PRISM Annual Review, Systems & Software track
Together with S. Fu (Lead Student) and F. Franchetti

2024 First Place, ACM Student Research Competition
The International Conference on Parallel Architectures and Compilation Techniques
Together with S. Fu (Lead Student) and F. Franchetti

2023 Outstanding Short Paper Award
IEEE High Performance Extreme Computing Conference
Together with P. Brinich, A. Ebel, F. Franchetti, and J. Johnson

2023 Second Place, ACM Student Research Competition
The International Symposium on Code Generation and Optimization
Together with F. Franchetti

2021 Discovery Scholar Distinction
University of Southern California

2018–21 Academic Achievement Award
University of Southern California

FELLOWSHIPS

2021–22 Carnegie Institute of Technology Dean's Fellowship

2019–21 University of Southern California Provost's Research Fellowship

GRANTS

2023- *High-Performance Code Generation for Homomorphic Encryption on GPUs using SPIRAL*
Tuned and benchmarked SPIRAL-generated number theoretic transform (NTT) implementations for homomorphic encryption (HE) applications on start-of-the-art GPUs.
N. Zhang (PI), F. Franchetti (Co-PI)
200,000 ACCESS Credits
NSF

RESEARCH EXPERIENCE

2025- *Durban: Enhancing Performance Portability in HPC Software with Artificial Intelligence*
Scaled up SPIRAL's semantics lifting capability via integration with neural code generation.
DoE

2023- *Code Synthesis for the PRISM Architecture*
Extended SPIRAL to target processing-in-memory (PIM) kernels on PRISM architectures.
SRC JUMP 2.0

2022- *Neocortex: SPIRAL Code Generation for Wafer-Scale Engine*
Extended SPIRAL to target Cerebras' second-generation Wafer-Scale Engine (WSE-2).
NSF

2024 *LLM Cerberus: Guarding LLMs against Hallucinating When Generating Mathematical Software*
Extended SPIRAL with symbolic execution and theorem proving to derive semantics and provide correctness guarantees for large language model (LLM)-generated math kernels.
NSF

2021-23 *Trebuchet: NTTX for OpenFHE*
Developed the SPIRAL NTTX package to automatically generate high-performance vectorized number theoretic transform (NTT) code for fully homomorphic encryption (FHE) applications.
DARPA DPRIVE

2020–21 *Compiler Abstractions Supporting High Performance on Extreme-scale Resources (CASPER)*
Developed a compiler-oriented autotuner that automatically profiles a kernel and performs tuning guided by performance prediction.
DARPA PAPPA

2019 *Dynamic Data-Aware Reconfiguration, INtegration and Generation (DDARING)*
Developed a lightweight augmented neural network for performance prediction.
DARPA SDH

PUBLICATIONS

Conference Proceedings

1. **N. Zhang**, S. McAleer, T. Sandholm. "Faster Game Solving via Hyperparameter Schedules." The AAAI Conference on Artificial Intelligence (AAAI), 2026. *To appear.*

2. Y. Lan, L. Tang, **N. Zhang**, Y. Eum, J. Hoe, F. Franchetti. "A RISC-V Vector Extension for Multi-word Arithmetic." The International Workshop on RISC-V for HPC (RISCV-HPC), in conjunction with the International Conference for High Performance Computing, Networking, Storage, and Analysis (SC), 2025.

3. **N. Zhang**, S. Fu, F. Franchetti. "Towards Closing the Performance Gap for Cryptographic Kernels Between CPUs and Specialized Hardware" The IEEE/ACM International Symposium on Microarchitecture (MICRO), 2025.

4. Q. Oschatz, **N. Zhang**, M. Franusich, F. Franchetti. "Towards Automated Reasoning Chains for Verification of LLM-Generated Scientific Code." IEEE High Performance Extreme Computing Conference (HPEC), 2025.

5. **N. Zhang**, F. Franchetti. "Code Generation for Cryptographic Kernels using Multi-word Modular Arithmetic on GPU." The International Symposium on Code Generation and Optimization (CGO), 2025.

6. **N. Zhang**, A. Ebel, N. Neda, P. Brinich, B. Reynwar, A. G. Schmidt, M. Franusich, J. Johnson, B. Reagen, F. Franchetti. "Generating High-Performance Number Theoretic Transform Implementations for Vector Architectures." IEEE High Performance Extreme Computing Conference (HPEC), 2023.

7. D. Sun, **N. Zhang**, F. Franchetti. "Optimization and Performance Analysis of Shor's Algorithm in Qiskit." IEEE High Performance Extreme Computing Conference (HPEC), 2023.

8. D. Soni, N. Neda, **N. Zhang**, B. Reynwar, H. Gamil, B. Heyman, M. N. T. Moopan, A. Al Badawi, Y. Polyakov, K. Canida, M. Pedram, M. Maniatakos, D. B. Cousins, F. Franchetti, M. French, A. Schmidt, B. Reagen. "RPU: The Ring Processing Unit." IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS), 2023.

9. **N. Zhang**, A. Srivastava, R. Kannan, V. K. Prasanna. "GenMAT: A General-Purpose Machine Learning-Driven Auto-Tuner for Heterogeneous Platforms." The Workshop on Programming Environments for Heterogeneous Computing (PEHC), in conjunction with the International Conference for High Performance Computing, Networking, Storage, and Analysis (SC), 2021.

10. A. Srivastava*, **N. Zhang***, R. Kannan, V. K. Prasanna. "Towards High Performance, Portability, and Productivity: Lightweight Augmented Neural Networks for Performance Prediction." The International Conference on High Performance Computing, Data, and Analytics (HiPC), 2020. **Equal contribution*.

11. C. Imes, A. Colin, **N. Zhang**, A. Srivastava, V. K. Prasanna, J. P. Walters. "Compiler Abstractions and Runtime for Extreme-scale SAR and CFD Workloads." The Workshop on Extreme Scale Programming Models and Middleware (ESPM2), in conjunction with the International Conference for High Performance Computing, Networking, Storage, and Analysis (SC), 2020.

Other Conference Papers, Technical Reports, Extended Abstracts, and Posters

1. **N. Zhang**, S. Rao, M. Franusich, F. Franchetti. "Towards Semantics Lifting for Scientific Computing: A Case Study on FFT." The Theory and Practice of Static Analysis Workshop (TPSA), in conjunction with the ACM SIGPLAN Symposium on Principles of Programming Languages (POPL), 2025, Extended abstract with presentation.

2. S. Fu, **N. Zhang**, F. Franchetti. "Accelerating High-Precision Number Theoretic Transforms using Intel AVX-512." The International Conference on Parallel Architectures and Compilation Techniques (PACT), 2024, Extended abstract with poster and presentation. **First Place, ACM Student Research Competition**. **Best Poster Runner-up** at PRISM Annual Review, Systems & Software track.

3. Y. Eum, **N. Zhang**, L. Tang, F. Franchetti. "Towards a RISC-V Instruction Set Extension for Multi-word Arithmetic." IEEE High Performance Extreme Computing Conference (HPEC), 2024, Extended abstract with poster.
4. P. Brinich, **N. Zhang**, A. Ebel, F. Franchetti, J. Johnson. "Twiddle Factor Generation for a Vectorized Number Theoretic Transform." IEEE High Performance Extreme Computing Conference (HPEC), 2023, Extended abstract with poster. **Outstanding Short Paper Award**.
5. H. Mankad, A. Rovinelli, M. Zecevic, P. McCorquodale, F. Franchetti, **N. Zhang**, S. Rao, R. A. Lebensohn, L. Capolungo. "EVPFFTX: A First Look at FFTX Applications in Material Science." IEEE High Performance Extreme Computing Conference (HPEC), 2023, Extended abstract with poster.
6. D. B. Cousins, Y. Polyakov, A. Al Badawi, M. French, A. Schmidt, A. Jacob, B. Reynwar, K. Canida, A. Jaiswal, C. Mathew, H. Gamil, N. Neda, D. Soni, M. Maniatakos, B. Reagen, **N. Zhang**, F. Franchetti, P. Brinich, J. Johnson, P. Broderick, M. Franusich, B. Zhang, Z. Cheng, M. Pedram. "TREBUCHET: Fully Homomorphic Encryption Accelerator for Deep Computation." The Government Microcircuit Applications and Critical Technology Conference (GOMACTech), 2023, Preprint with presentation.
7. **N. Zhang**, F. Franchetti. "Generating Number Theoretic Transforms for Multi-Word Integer Data Types." The International Symposium on Code Generation and Optimization (CGO), 2023, Extended abstract with poster and presentation. **Second Place, ACM Student Research Competition**.
8. **N. Zhang**, H. Gamil, P. Brinich, B. Reynwar, A. Al Badawi, N. Neda, D. Soni, K. Canida, Y. Polyakov, P. Broderick, M. Maniatakos, A. G. Schmidt, M. Franusich, J. Johnson, B. Reagen, D. B. Cousins, F. Franchetti. "Towards Full-Stack Acceleration for Fully Homomorphic Encryption." IEEE High Performance Extreme Computing Conference (HPEC), 2022, Extended abstract with presentation.
9. I. Grosof, **N. Zhang**, M. Heule. "Towards the shortest DRAT proof of the Pigeonhole Principle." The Pragmatics of SAT Workshop (PoS), in conjunction with the International Conference on Theory and Applications of Satisfiability Testing (SAT), 2022, Preprint with presentation.

TALKS

Seminars

2025 *Towards Closing the Performance Gap for Cryptographic Kernels Between CPUs and Specialized Hardware*
Computer Architecture Lab at Carnegie Mellon (CALCM), Oct. 2
Carnegie Mellon University, United States

2025 *Code Generation for Cryptographic Kernels using Multi-word Modular Arithmetic*
Ming Hsieh Department of Electrical and Computer Engineering, May 9
University of Southern California, United States

2025 *Code Generation for Cryptographic Kernels using Multi-word Modular Arithmetic*
Department of Electrical and Computer Engineering, May 2
New York University, United States

2025 *Optimization and Performance Analysis of Shor's Algorithm in Qiskit and Beyond*
The Center for Quantum Computing and Information Technologies (QCiT), Apr. 1
Carnegie Mellon University, United States

2025 *Code Generation for Cryptographic Kernels using Multi-word Modular Arithmetic*
The Programming Languages Group at the University of Pennsylvania (PLClub), Feb. 21
University of Pennsylvania, United States

2025 *Code Generation for Cryptographic Kernels using Multi-word Modular Arithmetic*
Computer Architecture Lab at Carnegie Mellon (CALCM), Feb. 14
Carnegie Mellon University, United States

Guest Lectures

2025 *Code Generation for Cryptographic Kernels using Multi-word Modular Arithmetic*
Computational Problem Solving for Engineers, Apr. 1
Carnegie Mellon University, United States

Conference and Workshop Presentations

2025 *Towards Closing the Performance Gap for Cryptographic Kernels Between CPUs and Specialized Hardware*
The IEEE/ACM International Symposium on Microarchitecture (MICRO), Oct. 22
Virtual

2025 *Code Generation for Cryptographic Kernels using Multi-word Modular Arithmetic*
The Workshop on Architectures for Zero-Knowledge Proofs and Verifiable Computation (ZKARCH), in conjunction with the IEEE/ACM International Symposium on Microarchitecture (MICRO), Oct. 18
Virtual

2025 *Towards Semantics Lifting for Scientific Computing: A Case Study on FFT*
Oak Ridge National Laboratory AI4Science Workshop, Apr. 30
Oak Ridge, United States

2025 *Code Generation for Cryptographic Kernels using Multi-word Modular Arithmetic on GPU*
The International Symposium on Code Generation and Optimization (CGO), Mar. 4
Las Vegas, United States

2025 *Towards Semantics Lifting for Scientific Computing: A Case Study on FFT*
The Theory and Practice of Static Analysis Workshop (TPSA), in conjunction with the ACM SIGPLAN Symposium on Principles of Programming Languages (POPL), Jan. 21
Denver, United States

2023 *Generating High-Performance Number Theoretic Transform Implementations for Vector Architectures*
IEEE High Performance Extreme Computing Conference (HPEC), Sep. 29
Virtual

2023 *Generating Number Theoretic Transforms for Multi-Word Integer Data Types*
The International Symposium on Code Generation and Optimization (CGO), Feb. 28
Montreal, Canada

2022 *Towards Full-Stack Acceleration for Fully Homomorphic Encryption*
IEEE High Performance Extreme Computing Conference (HPEC), Sep. 23
Virtual

2021 *GenMAT: A General-Purpose Machine Learning-Driven Auto-Tuner for Heterogeneous Platforms*
The Workshop on Programming Environments for Heterogeneous Computing (PEHC), in conjunction with the International Conference for High Performance Computing, Networking, Storage, and Analysis (SC), Nov. 19
Virtual

2020 *Towards High Performance, Portability, and Productivity: Lightweight Augmented Neural Networks for Performance Prediction*
The International Conference on High Performance Computing, Data, and Analytics (HiPC), Dec. 16
Virtual

Tutorials

2026 *SPIRAL: Pre-Silicon and Early-Prototype Performance Estimation Using Highly Optimized Code*
ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS), *To appear*
Together with F. Franchetti
Pittsburgh, United States

2025 *Open Source SPIRAL 8.5.1 Tutorial*
IEEE High Performance Extreme Computing Conference (HPEC), Sep. 17
Together with F. Franchetti and M. Franusich
Virtual

2024 *Open Source SPIRAL 8.5 Tutorial*
IEEE High Performance Extreme Computing Conference (HPEC), Sep. 25
Together with F. Franchetti and M. Franusich
Virtual

2023 *Open Source SPIRAL 8.5 Tutorial*
IEEE High Performance Extreme Computing Conference (HPEC), Sep. 27
Together with F. Franchetti, M. Franusich, and P. Broderick
Virtual

TEACHING EXPERIENCE

Carnegie Mellon University

Teaching Assistant

24 Fall	Mathematical Foundations of Electrical Engineering
23 Spring	Computational Problem Solving for Engineers

University of Southern California

Undergraduate Teaching Assistant

21 Spring	Special Topics - Accelerated Computing Using FPGAs
20 Fall	Parallel and Distributed Computation
20 Spring	Special Topics - Accelerated Computing Using FPGAs
20 Spring	Discrete Methods in Computer Science
19 Fall	Parallel and Distributed Computation
19 Fall	Discrete Methods in Computer Science

MENTORING

Undergraduate

2024-	Misho Alexandrov
2024-	Sophia Fu
2023-	Gordon Xu
2024	Govind Malasani
2025	Yiwen Jiang
2024	Zubin Narayan
2024	Youngjin Eum
2024	Steven Lee
2022–23	Matt Ngaw
2022–23	Jimmy Zhou

Master’s

2025-	Yunhao Lan
2024-25	Yujun Lee
2023	Kofi Poku
2022–23	Dewang Sun
2022	Hongbo Sun

SERVICE

Conference Program Committees

The AAAI Conference on Artificial Intelligence (AAAI), 2026

The Workshop on AI Assisted Software Development for HPC (AI4Dev), in conjunction with the International Conference on Parallel Processing (ICPP), 2025

Journal Peer Review

IEEE Transactions on Dependable and Secure Computing (TDSC)

IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD)

IEEE Transactions on Computational Social Systems (TCSS)

IEEE Transactions on Mobile Computing (TMC)

IEEE Transactions on Emerging Topics in Computing (TETC)

ACM Computing Surveys (CSUR)

The International Journal of High Performance Computing Applications (IJHPCA)

IEEE Transactions on Parallel and Distributed Systems (TPDS)

IEEE Transactions on Information Forensics & Security (T-IFS)

Service to the University

Carnegie Institute of Technology College Council, 2025-26

CMU Electrical and Computer Engineering Faculty Hiring Student Council, 2022-25

Outreach

CMU College of Engineering Graduate Student Outreach Committee, 2023

Updated December 2025