Xuesi Chen

xc562@cornell.edu | xuesichen.github.io | github.com/XuesiChen

Research Interests

My research focuses on designing energy-efficient computing systems from a full-stack perspective, spanning compiler optimization to embedded edge platforms. I develop models and tools to understand the relationships between performance, energy use, and environmental impact, and explore hardware–software co-design strategies that improve efficiency across the system stack. Much of my recent work looks at edge devices, where I study how design choices affect both system performance and long-term sustainability.

Education

Cornell Tech 2025 – present

- PhD in Electrical and Computer Engineering, GPA: 4.3/4.3
- Advisor: Udit Gupta

Carnegie Mellon University 2022 – 2024

- M.S. in Electrical and Computer Engineering, GPA: 3.85/4.0
- Advisors: Brandon Lucia and Nathan Beckmann

Tufts University 2018 – 2022

- B.S. in Computer Engineering, GPA: 3.78/4.0
- Advisor: Mark Hempstead

Relevant Coursework: Computer Architecture, Computer Systems, Parallel Computing, Reconfigurable Logic, Optimizing Compiler, Emerging Memory Technologies, VLSI Design, Computer Networks, Machine Learning Hardware and Systems.

Professional Experience

Research Assistant, Cornell Tech – New York, NY June 2024 – Present

Advisor: Udit Gupta

- Characterized the performance–energy trade-offs of running LLMs on workbench-scale GPUs and evaluated model accuracy to identify optimal strategies for deploying LLMs under the memory capacity and energy constraints of edge hardware.
- Designing a system-level framework for developing energy-efficient and sustainable IoT devices by jointly modeling performance, energy, and carbon impacts of off-the-shelf microcontrollers running TinyML models under varying deployment and application constraints (preliminary work published at HotCarbon 2025, a full conference paper to be appeared at MobiSys 2026).
- Prototying a camera-based visitor-counting edge device using YOLOv8 and ByteTrack tracking on the Coral Dev Board Micro to help New York City Parks monitor visitor flows at park entrances at scale.
- Developed a probabilistic modeling framework to capture uncertainty in embodied carbon and studied its implications for performance trade-offs and chiplet designs (published at ICCAD 2025).
- Analyzing trade-offs between latency, user experience, and carbon impact in LLM serving, incorporating user preferences to balance service-level objectives (SLOs) with sustainability goals (published at JCSS 2025).

Research Assistant, Carnegie Mellon University – Pittsburgh, PA Sept 2022 – May 2024

- Developed a compiler and hardware co-design solution for time-multiplexing PEs on energy-minimal CGRAs for PE utilization enhancement (published at YArch 2023).
- Engineered a compiler to efficiently schedule and map instructions to handle extreme-edge computing workloads with minimized energy consumption.
- Conducted performance, energy and area analysis using the PE-level RTL synthesis of proposed hardware designs and simulation-based events counting.
- Solved the challenge of executing large workloads on energy-minimal edge processors by improving utilization by 2x with only 5% energy overhead compared to the state-of-the-art RipTide architecture.

Undergraduate Indepedent Researcher, Tufts University – Medford, MA Oct 2020 – May 2022

- Verified the LLC behavior of a multi-core cache contention simulator PInTE and compared the real contention results with the simulator contention results using RMSE and KL Divergence (published at IISWC 2022).
- Designed and implemented an thermal hotspots simulator by extend and integrate the compute activity simulation based on SCALE-SIM with processor hotspot categorizer HotGauge for neural network accelerators (published at HSSB 2022).

Software Engineer Intern, Amazon – Cambridge, MA Summer 2021

- Implemented and tested a workflow that terminates edge-to-cloud connection for Alexa upon receiving a false awake signal.
- Implemented workflow result handlers that store and transmit metadata and metrics related to all incoming wake word initiations.
- Launched dashboards on AWS CloudWatch to track all metrics corresponding to incoming wake word invocations for research and workflow health monitoring purposes.

Peer-reviewed Publications

Enabling Carbon-aware Edge System Design MobiSys Rasing Star 2026

Xuesi Chen

A Greener Edge: A Framework on Carbon-aware Edge ML System Design MobiSys 2026

Xuesi Chen, Ilan Mandel, Eren Yildiz, Josiah Hester, Udit Gupta

When the LLM Slows Down: How Environmental Impacts Mediate University Students' LLM Interactions ICT4S 2026

Hyeonwook Kim, **Xuesi Chen**, Alex Cabral, Cindy Kaiying Lin, Josiah Hester, Udit Gupta

COFFEE: A Carbon-Modeling and Optimization Framework for HZO-based FeFET eNVMs DATE 2026

Hongbang Wu, **Xuesi Chen**, Shubham Jadhav, Amit Lal, Lillian Pentecost, Udit Gupta

CarbonClarity: Understanding and Addressing Uncertainty in Embodied Carbon for Sustainable Computing ICCAD 2025

Xuesi Chen, Leo Han, Anvita Bhagavathula, Udit Gupta

Slower is Greener: Acceptance of Eco-feedback Interventions on Carbon Heavy Internet Services JCSS 2025

Haisley Kim, Sydney Young, **Xuesi Chen**, Udit Gupta, Josiah Hester

PInTE: Probabilistic Induction of Theft Evictions IISWC 2022

Cesar Gomes, **Xuesi Chen**, Mark Hempstead

Workshop Publications

From Component to System: Rethinking Edge Computing Design Through a Carbon-Aware Lens HotCarbon 2025

Xuesi Chen, Ariel Goldner, Eren Yildiz, Ilan Mandel, Tingyu Cheng, Josiah Hester, Udit Gupta

Dataflow Blocks: Modular Time-Multiplexing for CGRAs YArch 2023 (@ ASPLOS)

Xuesi Chen, Nishanth Subramanian, Karthik Ramanathan, Nathan Beckmann, Brandon Lucia

NNShim: Thermal Hotspots Simulation on ML Accelerators HSSB 2022 (@ ISCA)

Xuesi Chen, Daniel Ernst, Margret Riegert, Mark Hempstead

Designing Equitable Scheduling Systems CWIDCA 2022 (@ MICRO)

Sahana Rangarajan, **Xuesi Chen**, Pratyush Patel, Sara Mahdizadeh Shahri, Jaylen Wang, Akshitha Sriraman

Awards and Honors

MICRO WICArch Early-Career Fellow 2024

Organized and led a gathering of 40+ participants for Women in Computer Architecture at MICRO

Harry Poole Burden Prize 2022
Best research project by ECE undergraduates

Teaching

Cornell EE6950: Architecture for AI Computing Systems Fall 2025
Coordinated course logistics and advised student projects
Tufts EE156: Advanced Computer Architecture Spring 2022
Graded homework and labs, hosted office hours, and advised student projects

Skills

Programming Languages: Python, C/C++, Java, Shell, VHDL, Verilog, Assembly, CUDA, MATLAB
Tools and Simulators: LLVM, ModelSim, SCALE-Sim, McPAT
Embedded Systems: Raspberry Pi, Arduino, Coral edgeTPU
Sustainability: LCA analysis for electronic components