

Udit Gupta

Incoming Assistant Professor in ECE – Cornell Tech
✉ ugupta@g.harvard.edu • 📩 ugupta.com

RESEARCH OVERVIEW

My research spans across computer architecture, systems, and machine learning to co-design solutions across the computing stack. Generally, I am interested in discovering and demonstrating new ways to design systems and hardware to improve the performance, efficiency, and environmental sustainability of emerging applications.

EDUCATION

Harvard University, Ph.D. Cambridge, MA
Computer Science 2016-November 2022

Advisors: Professor David Brooks and Professor Gu-Yeon Wei

Dissertation: Enabling High Performance, Efficient, and Sustainable Deep Learning Systems At Scale

Cornell University, B.S. Ithaca, NY
Electrical & Computer Engineering, Computer Science 2012-2016

GPA: 4.0, *summa cum laude*

Advisor: Professor Zhiru Zhang

Undergraduate research: Using C-to-gates EDA tools like high-level synthesis to build hardware accelerators

PROFESSIONAL EXPERIENCE

Cornell Tech New York City NY
Assistant Professor, Electrical and Computer Engineering Summer 2023-

Facebook AI Research (FAIR) Menlo Park, CA
Visiting Research Scientist October 2018-May 2023

Advisors: Dr. Carole-Jean Wu and Dr. Hsien-Hsin S. Lee

- Developed architectural models to enable life cycle carbon accounting for real mobile and datacenter scale servers (published at ISCA 2022 and IEEE MICRO Top Picks). Collaborated with sustainability and infrastructure teams to develop a datacenter wide total-carbon-ownership dashboard.
- Investigated Meta's end-to-end AI carbon footprint, demonstrating holistic mitigation strategies are needed across hardware and model life cycles for sustainable AI (published at ISCA 2022, MLSys 2023, and IEEE MICRO Top Picks).
- Designed specialized hardware accelerators for efficiently serving at-scale neural recommendation models run-time schedulers to optimize (published at MICRO 2021).
- Implemented run-time schedulers to optimize recommendation inference across diverse models and hardware. Optimizations reduced tail-latency by 40% in production use cases (published at ISCA 2020).
- Led the first characterization of Facebook's production recommendation models, outlining paths for future AI hardware design and academic research into systems for recommendation (published at HPCA 2020).

Algo-Logic Systems Santa Clara, CA
Hardware Design and Verification Engineering Intern Summer 2015

- Designed and implemented OpenCL software kernels for financial data parsers on FPGAs.

HONORS AND AWARDS

- IEEE MICRO Top Picks 2023
Top 12 across all papers published at computer architecture venues in 2022 recognized.

- IEEE MICRO Top Picks 2022
Top 12 across all papers published at computer architecture venues in 2021 recognized.
- IEEE MICRO Top Picks Honorable Mention 2021
Top 24 across all papers published at computer architecture venues in 2020 recognized.
- Best Paper Nominee at Parallel Architectures and Compilation Techniques (PACT) 2019
Top 4 of 38 papers at PACT conference.
- Best Paper Nominee at Design Automation Conference (DAC) 2018
Top 16 of 178 papers presented at DAC nominated.
- Harvard Smith Family Fellowship 2017-2018
Awarded fellowship for 1 year of tuition and stipend (\$80K), plus additional \$5K of research funds.
- National Science Foundation (NSF) GRFP Honorable Mention 2016
- Richard A. Newton Young Fellows Scholarship at DAC 2015 2015
Fellowship for early career researchers to attend DAC. Award included complementary registration and travel funds.
- Cornell Eta Kappa Nu (HKN) - Electrical Engineering Honor Society 2013 - 2016
- Cornell ECE Early Research Career Scholarship 2013
Scholarship to conduct undergraduate summer research with computer systems lab. Award included \$4000 stipend.

PATENTS

Computer memory module processing device with cache storage

Liu Ke, Xuan Zhang, **Udit Gupta**, Carole-Jean Wu, Mark David Hempstead, Brandon Reagen, Hsien-Hsin Sean Lee

US11442866B2

PEER-REVIEWED PUBLICATIONS

MP-Rec: Hardware-Software Co-Design to Enable Multi-Path Recommendation

Samuel Hsia, **Udit Gupta**, Bilge Acun, Newsha Ardalani, Pan Zhong, Gu-Yeon Wei, David Brooks, Carole-Jean Wu

To appear in ASPLOS 2023 (available on [arXiv](#)).

Co-designing embedding representations and hardware platforms to improve algorithmic and systems performance.

A Holistic Approach for Designing Carbon Aware Datacenters

Bilge Acun, Benjamin Lee, Kiwan Maeng, Manoj Chakkaravarthy, **Udit Gupta**, David Brooks, Carole-Jean Wu
To appear in ASPLOS 2023 (available on [arXiv](#)).

A tool to balance 24/7 renewable energy, energy storage, and workload shifting to optimize operational and embodied carbon.

ACT: Designing Sustainable Computer Systems With An Architectural Carbon Modeling Tool

Udit Gupta, Mariam Elgamal, Gage Hills, Gu-Yeon Wei, Hsien-Hsin S. Lee, David Brooks, Carole-Jean Wu

The 49th IEEE/ACM International Symposium on Computer Architecture (ISCA 2022).

An architectural carbon modeling tool to quantify the carbon footprint from hardware manufacturing.

IEEE MICRO Top Picks 2023

Sustainable AI: Environmental Implications, Challenges and Opportunities

Carole-Jean Wu, Ramya Raghavendra, **Udit Gupta**, Bilge Acun, Newsha Ardalani, Kiwan Maeng, Gloria Chang, Fiona Aga Behram, James Huang, Charles Bai, Michael Gschwind, Anurag Gupta, Myle Ott, Anastasia Melnikov, Salvatore Candido, David Brooks, Geeta Chauhan, Benjamin Lee, Hsien-Hsin S Lee, Bugra Akyildiz, Maximilian

Balandat, Joe Spisak, Ravi Jain, Mike Rabbat, Kim Hazelwood
Machine Learning and Systems (MLSys 2022, available on [arXiv](#)).

An industry perspective on quantifying and understanding the environmental implications of AI at-scale.

Hercules: Heterogeneity-Aware Inference Serving for At-scale Personalized Recommendation

Liu Ke, Udit Gupta, Mark Hempstead, Carole-Jean Wu, Hsien-Hsin Sean Lee, Xuan Zhang

IEEE International Symposium on High-Performance Computer Architecture (HPCA 2022).

Optimizing recommendation model scheduling and placement across heterogeneous commodity and specialized hardware.

RecPipe: Co-designing Models & Hardware to Jointly Optimize Recommendation Quality & Performance

Udit Gupta, Samuel Hsia, Jeff Zhang, Mark Wilkering, Javin Pombra, Hsien-Hsin S. Lee, Gu-Yeon Wei, Carole-Jean Wu, David Brooks

IEEE/ACM International Symposium on Microarchitecture (MICRO 2021).

A system to optimize multi-stage recommendation pipelines using specialized accelerators.

Artifact Badges: Available, Functional, and Reproducible

Code available on [GitHub](#)

RecSSD: Near Data Processing for Solid State Drive Based Recommendation Inference

Mark Wilkering, **Udit Gupta**, Samuel Hsia, Caroline Trippel, Carole-Jean Wu, David Brooks, Gu-Yeon Wei

Int. Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS 2021)

Optimizing inference latency with near data processing on SSD's for large-scale recommendation models.

Artifact Badges: Available

Chasing Carbon: The Elusive Environmental Footprint of Computing

Udit Gupta, Young Geun Kim, Sylvia Lee, Jordan Tse, Hsien-Hsin S. Lee, Gu-Yeon Wei, David Brooks, Carole-Jean Wu

IEEE International Symposium on High-Performance Computer Architecture (HPCA 2021)

Characterizes the carbon footprint of mobile and data center-scale systems across hardware life cycles including manufacturing and operational use.

IEEE MICRO Top Picks 2022

Featured by: [Harvard Gazette](#), [Tech at Facebook](#), [Bloomberg Green](#), [The Guardian](#) [CNBC](#)

Cross-Stack Workload Characterization of Deep Recommendation Systems

Samuel Hsia, **Udit Gupta**, Mark Wilkering, Carole-Jean Wu, Gu-Yeon Wei, David Brooks

IEEE International Symposium on Workload Characterization (IISWC 2020)

Characterizes recommendation workloads using Intel TopDown performance analysis method.

DeepRecSys: A System for Optimizing End-To-End At-scale Neural Recommendation Inference

Udit Gupta, Samuel Hsia, Vikram Saraph, Xiaodong Wang, Brandon Reagen, Gu-Yeon Wei, Hsien-Hsin S. Lee, David Brooks, Carole-Jean Wu

The 47th IEEE/ACM International Symposium on Computer Architecture (ISCA 2020).

A system to optimize latency-bounded throughput of recommendation engines across varying loads, network architecture, and heterogeneous hardware.

Code available on [GitHub](#).

Load generation characteristics merged with standardized MLPerf inference benchmark.

Scheduling optimizations reduced tail-latency by 40% in Facebook's production recommendation use cases.

RecNMP: Accelerating Personalized Recommendation with Near-Memory Processing

Liu Ke, **Udit Gupta**, Carole-Jean Wu, Benjamin Youngjae Cho, Mark Hempstead, Brandon Reagen, Xuan Zhang, David Brooks, Vikas Chandra, Utku Diril, Amin Firoozshahian, Kim Hazelwood, Bill Jia, Hsien-Hsin S. Lee, Meng Li, Bert Maher, Dheevatsa Mudigere, Maxim Naumov, Martin Schatz, Mikhail Smelyanskiy, Xiaodong

Wang

The 47th IEEE/ACM International Symposium on Computer Architecture (ISCA 2020).

Accelerating recommendation inference with near memory processing augmented DRAM.

Architectural Implications of Facebook's DNN-based Personalized Recommendation

Udit Gupta, Xiaodong Wang, Maxim Naumov, Carole-Jean Wu, Brandon Reagen, David Brooks, Bradford Cottel, Kim Hazelwood, Bill Jia, Hsien-Hsin S. Lee, Andrey Malevich, Dheevatsa Mudigere, Mikhail Smelyanskiy, Liang Xiong, Xuan Zhang

IEEE International Symposium on High-Performance Computer Architecture (HPCA 2020)

Quantifies the architectural implications of neural personalized recommendation models and charts paths for future hardware optimization to accelerate recommendation inference.

IEEE MICRO Top Picks 2021 - Honorable Mention

Featured by: [Facebook Research](#)

MASR: A Modular Accelerator for Sparse RNNs

Udit Gupta, Brandon Reagen, Lillian Pentecost, Marco Donato, Thierry Tambe, Alexander Rush, Gu-Yeon Wei, David Brooks

Parallel Architectures and Compilation Techniques (PACT 2019).

Designs a sparse recurrent neural network accelerator based on a low-cost, compute intensive sparse encoding technique.

Best Paper Nominee

MaxNVM: Maximizing DNN Storage Density and Inference Efficiency with Sparse Encoding and Error Mitigation

Lillian Pentecost, Marco Donato, Brandon Reagen, **Udit Gupta**, Siming Ma, Gu-Yeon Wei, David Brooks.

IEEE/ACM International Symposium on Microarchitecture (MICRO 2019).

A framework to maximize DNN inference efficiency by using on-chip embedded non-volatile memories.

A 16nm 25mm² SoC with a 54.5× Flexibility-Efficiency Range from Dual-Core Arm Cortex-A53, to eFPGA, and Cache-Coherent Accelerators

Paul Whatmough, Sae Kyu Lee, Marco Donato, Hsea-Ching Hseuh, Sam Xi, **Udit Gupta**, Lillian Pentecost, Glenn Ko, David Brooks, Gu-Yeon Wei.

Symposia on VLSI Technology and Circuits. (VLSI 2019)

A state-of-the-art SoC with mobile CPUs, embedded FPGA, and cache-coherent neural network accelerators.

SMIV: A 16nm SoC with Efficient and Flexible DNN Acceleration for Intelligent IoT Devices.

Paul Whatmough, Sae Kyu Lee, Sam Xi, **Udit Gupta**, Lillian Pentecost, Marco Donato, Hsea-Ching Hseuh, David Brooks, Gu-Yeon Wei.

Hot Chips (Hot Chips 2018).

A specialized and flexible SoC for efficient DNNs in 16nm technology.

Weightless: Lossy Weight Encoding for Deep Neural Network Compression.

Brandon Reagen, **Udit Gupta**, Robert Adolf, Michael Mitzenmacher, Alexander Rush, Gu-Yeon Wei, David Brooks.

International Conference on Machine Learning (ICML 2018).

A lossy weight compression technique using Bloomier filters to compress deep neural networks for over the wire compression.

Ares: A Framework for Quantifying the Resilience of Deep Neural Networks.

Brandon Reagen, **Udit Gupta**, Lillian Pentecost, Paul Whatmough, Sae Kyu Lee, Niamh Mulholland, Gu-Yeon Wei, David Brooks.

Design Automation Conference (DAC 2018).

A Python-based tool to quantify the bit-level fault tolerance and resilience of deep neural networks.

Best Paper Nominee

On-chip Deep Neural Network Storage with Multi-level eNVM.

Marco Donato, Brandon Reagen, Lillian Pentecost, **Udit Gupta**, David Brooks, Gu-Yeon Wei.

Design Automation Conference (DAC 2018).

A tool to quantify the performance, storage, and energy efficiency improvements of multi-level embedded non-volatile memories for neural networks.

Rosetta: A Realistic Benchmark Suite for Software Programmable FPGAs.

Yuan Zhou, **Udit Gupta**, Steve Dai, Ritchie Zhao, Nitish Srivastava, Hanchen Jin, Joseph Featherston, Yi-Hsiang Lai, Gai Liu, Gustavo Velasquez, Wenping Wang, Zhiru Zhang.

ACM/SIGDA International Symposium on Field-Programmable Gate Arrays (FPGA 2018)

A benchmark suite for high-level synthesis-based FPGA acceleration.

Dynamic Hazard Resolution for Pipelining Irregular Loops in High-Level Synthesis.

Steve Dai, Ritchie Zhao, Gai Liu, Shreesha Srinath, **Udit Gupta**, Christopher Batten, Zhiru Zhang.

ACM/SIGDA International Symposium on Field-Programmable Gate Arrays (FPGA 2017)

Improving performance of high-level synthesis-based accelerator designs by resolving memory hazards for pipelines with irregular loops.

Mapping-Aware Constrained Scheduling for LUT-Based FPGAs.

Mingxing Tan, Steve Dai, **Udit Gupta**, Zhiru Zhang.

ACM/SIGDA International Symposium on Field-Programmable Gate Arrays (FPGA 2015)

Optimizing resource scheduling for high-level synthesis based FPGA designs.

WORKSHOP PUBLICATIONS

Mentoring Opportunities in Computer Architecture: Analyzing the Past to Develop the Future

Elba Garza, Gururaj Saileshwar, **Udit Gupta**, Tianyi Liu, Abdulrahman Mahmoud, Saugata Ghose, Joel Emer
Workshop on Computer Architecture Education (WCAE) in conjunction with ISCA 2021

Outlined the state of mentoring programs in computer architecture and charted future paths from CASA student group.

Quantifying the impact of data encoding on DNN fault tolerance

Edward Pyne, Lillian Pentecost, **Udit Gupta**, Gu-Yeon Wei, David Brooks

Workshop on Performance Analysis of Machine Learning Systems (FastPath) workshop at ISPASS 2020

Quantified the impact of data encoding on DNN fault tolerance and resilience.

MASR: Modular Accelerator for Sparse RNNs

Udit Gupta, Brandon Reagen, Lillian Pentecost, Marco Donato, Thierry Tambe, Alexander Rush, Gu-Yeon Wei, David Brooks

Cognitive Architectures (CogArch) workshop at HPCA 2020

Designs a sparse recurrent neural network accelerator based on a low-cost, compute intensive sparse encoding technique.

Human Activity Recognition Using Wearables and a Low-power Deep Neural Network Accelerator

Sreela Kodali, **Udit Gupta**, Lillian Pentecost, David Brooks, Gu-Yeon Wei

SRC TechCon workshop 2017

A system for human-activity recognition for embedded, wearable devices using deep neural network accelerators.

TECHNICAL ARTICLES

I am eager to build broader research communities to build efficient, scalable, and sustainable hardware. I often write technical articles to help distill key research challenges and opportunities for broader audiences.

Designing low-carbon computers with an architectural carbon modeling tool

Carole-Jean Wu, Udit Gupta

Tech at Facebook, June 2022

Technical blog outlining our efforts to build models to predict computing hardware's carbon footprint.

Most of computing's carbon emissions are coming from manufacturing and infrastructure

Carole-Jean Wu, Udit Gupta

Tech at Facebook, March 2021

Technical blog summarizing Facebook's carbon footprint based on Chasing Carbon paper (HPCA 2021)

Optimizing Infrastructure for Neural Recommendation At-scale

Carole-Jean Wu, Udit Gupta

Facebook AI, February 2020

Technical blog outlining new hardware optimization paths for personalized recommendation based on HPCA 2020 paper.

Deep Learning: It's Not All About Recognizing Cats and Dogs

Carole-Jean Wu, David Brooks, Udit Gupta , Hsien-Hsin Lee, and Kim Hazelwood

ACM SIGARCH, Computer Architecture Today, November 2019

Call to action for computer architects to design hardware for DNN-based recommendation models. I surveyed the breakdown of types of DNNs optimized across computer architecture papers to show underinvestment in personalized recommendation.

Designing AI-Enabled Technology for Society

Udit Gupta, Lillian Pentecost

Harvard Science in the News (SITN), October 2018

Co-developed and gave lecture on the societal impact of AI for general public.

Software Programmable FPGAs

Udit Gupta

Circuit Cellar (Tech the Future series), June 2017

Technical article on more productive, C-to-gates programming tools for FPGAs based on undergraduate research.

PRESS

The global chip industry has a colossal problem with carbon emissions

CNBC, November 2021

Highlights the rising carbon footprint of chip manufacturing based on our Chasing Carbon (HPCA 2021) paper.

The Computer Chip Industry Has a Dirty Climate Secret

The Guardian, September 2021

Highlights the rising carbon footprint of semi-conductor manufacturers based on our Chasing Carbon (HPCA 2021) paper.

The Chip Industry Has a Problem With Its Giant Carbon Footprint

Bloomberg, April 2021

Highlights the rising carbon footprint of semi-conductor manufacturers based on our Chasing Carbon (HPCA 2021) paper.

Smaller, Faster, Greener

Harvard Gazette, March 2021

Summarizes our Chasing Carbon paper (HPCA 2021) on computing's carbon footprint across hardware life cycles.

Facebook Open Sourced this Architecture for Personalized Neural Recommendation Systems

Medium (Data Series), May 2020

Summarizes our hardware performance analysis (HPCA 2020) of Facebook's production recommendation models.

OPEN SOURCE TOOLS AND INFRASTRUCTURE

I strive to open-source tools and resources to catalyze academic systems and architecture research in new areas. My efforts have focused on building accessible tools for systems research in recommendation engines.

Carbon Explorer: A Holistic Approach for Designing Carbon Aware Datacenters

A tool to evaluate solutions that make datacenters operate on renewable energy holistically by including embodied and operational footprints. .

Code available on [GitHub](#)

ACT: Architectural Carbon Modeling Tool

A model to estimate the carbon footprint of computing hardware across life cycles including manufacturing and use.

Code available on [GitHub](#)

RecPipe: Co-designing Models and Hardware to Optimize Recommendation Quality and Performance

A framework to study multi-stage recommendation on commodity hardware and simulated accelerators (MICRO 2021).

Code available on [GitHub](#)

Received Artifact available, functional, and reproducible badges at MICRO 2021

DeepRecSys: A System for Optimizing End-To-End At-scale Neural Recommendation Inference

Infrastructure to optimize latency-bounded throughput for recommendation workloads (ISCA 2020).

Code available on [GitHub](#)

Recommendation load generator characteristics and insights merged with industry-standard MLPerf benchmark

Deep Learning Recommendation Model (DLRM) for Personalization and Recommendation Systems

Facebook's open-source deep learning recommendation model in PyTorch and Caffe2 ([arXiv](#)).

Code available on [GitHub](#)

Facebook's DLRM adopted by industry-standard MLPerf training and inference benchmarks

Ares: A framework for quantifying the resilience of deep neural networks

Tool to quantify the fault tolerance and resilience of deep neural networks (DAC 2018).

Code available on [GitHub](#)

MLPerf: A Benchmark for Machine Learning from an Academic/Industry Cooperative

Industry-academic standardized benchmark for machine learning system performance

Code available on [MLPerf](#)

Helped build the initial MLPerf benchmark and guide recommendation inference benchmark design

PROFESSIONAL SERVICE AND COMMUNITY INVOLVEMENT

Computer Architecture Student Association (CASA)

Steering Committee Member 2020-2021

Co-Chair 2022-Present

- Student group fostering an inclusive computer architecture community.
- Organized [reading group](#) to advance and promote diversity, equity, and inclusion in computer architecture.
- Gathered qualitative and quantitative feedback on mentoring and networking initiatives in computer architecture, published at [Workshop on Computer Architecture Education \(WCAE\)](#) at ISCA 2021.

Conference Organizing Committees

- MLSys Organizer as “Panels and Young Activities Chair” MLSys 2022, MLSys 2023
Organizing activities to grow the presence of junior researchers in the machine learning and systems community.

Technical Workshop and Tutorial Organizing Committees

- NetZero Carbon co-founder and co-organizer HPCA 2023
Venue for developing zero-carbon computing platforms with over 25 attendees, 2 keynote speakers, and 14 submitted papers.
- PeRSonAl co-founder and co-organizer ASPLOS 2020, ISCA 2020, MLSys 2021
Venue for designing personalized recommendation engines across application, algorithms, and systems and hardware with over 200 attendees over 50 institutions.
- CLEAR co-founder and co-organizer ISCA 2021
Venue for Computing Landscapes for Environmental Accountability and Responsibility (CLEAR) with over 50 attendees, 9 invited talks, and industry-academic panels.
- JOURNE co-founder and co-organizer MLSys 2021
Venue for share the evolution of research ideas through specific examples of negative results, retrospectives, and project post-mortems.
- NOPE co-organizer ASPLOS 2019, ASPLOS 2021
Venue for sharing negative outcomes, post-mortems, and experience in research.

Conference Program Committees

- IEEE MICRO (MICRO) 2023 program committee member
- Machine Learning and Systems (MLSys) 2023 program committee member
- Neural Information Processing Systems (*NeurIPS*) 2022 reviewer
- International Conference on Learning Representations (*ICLR*) 2019 reviewer

External/Sub-session/Workshop Program Committees

- HotCarbon 2023 (co-located with OSDI) program committee member
- Nature Communications 2023 reviewer
- MICRO 2022 student research competition (SRC) reviewer
- PACT 2022 student research competition (SRC) reviewer

Professional Memberships

- 3C (Cultural Competence in Computing) Fellow 2021-2022
- Harvard SITN Lecture Series , Lecture Facilitator and Director 2018-2019
- Cornell IEEE Student Chapter, President and Corporate Director 2013-2016

TEACHING EXPERIENCE

- Graduate Teaching Fellow** Cambridge, MA
 Harvard University
- CS 290: PhD Grad Cohort Research Seminar Fall 2020
Head teaching fellow for research seminar course for incoming CS PhD students. Organized readings and invited faculty talks across a variety of CS research areas. Lead recitation sections.
 - CS 141: Computing Hardware Spring 2019
Designed homework assignments and exams for introductory computer engineering course. Lead office hours and recitation sections.

- Undergraduate Teaching Assistant** Ithaca, NY
 Cornell University
 - CS 3420/ECE 3140: Embedding Systems Spring 2016
Mentored final embedded systems projects, held office hours, and graded assignments.
 - ECE 2300: Introduction to Digital Logic and Computer Organization Fall 2015, Spring 2015, Spring 2014
Hosted multiple weekly lab sessions and office hours for introductory course in computer engineering.

- edX Course Development** Ithaca, NY
 Cornell University Summer 2014
 - Designed lab and quizzes for EdX MOOC course “*The Computing Inside Your Smartphone*”.
 - Translated assignments from introduction to computer engineering course for online format and audience.

Education Outreach for Middle School and High School Students

- Designed a short course for middle school and high school students on sustainable computing to highlight the societal impact of computing. Taught the course in *Summer 2021 and Spring 2021* to 50 students at *MIT Splash!*.
- Designed a lab-driven short course for high school students on computer engineering. Taught the course in *Fall 2014 and Spring 2015* to 30 students at *Cornell Splash!*. After the course many students asked where to buy circuit components to use at home.

RESEARCH MENTORING

Mariam Elgamal (PhD Student) August 2021 - Present
Understanding and optimizing the carbon footprint of hardware manufacturing.

Samuel Hsia (PhD Student) August 2019 - Present
Characterizing hardware performance of neural recommendation models using Intel TopDown (IISWC 2020)

Liu Ke (PhD student) May 2019 - January 2021
Designing near memory processing accelerators for data center scale personalized recommendation (RecNMP, ISCA 2020)

Jaylen Wang (4th year undergraduate at Harvard University) August 2021 - September 2021
Evaluating carbon footprint of cryptocurrency mining systems across hardware life cycles.

Javin Pombra (3rd year undergraduate at Harvard University) May 2020 - September 2021
Evaluating trade offs between fairness and accuracy for recommendation models.
Designing neural recommendation models for multi-stage recommendation (co-author on RecPipe, MICRO 2021).

Lucy He (2nd year undergraduate at Harvard University) May 2020 - September 2020
Designing and building a training model zoo for neural recommendation.

Tarun Prasad (2nd year undergraduate at Harvard University) May 2020 - September 2020
Designing and building a training model zoo for neural recommendation.

Festus Ojo (3rd year undergraduate at Harvard University) May 2020 - September 2020
Quantifying the performance and accuracy tradeoffs of probabilistic recommendation models.

Ted Pyne (3rd year undergraduate at Harvard University) January 2020 - August 2020
Designing methods to improve fault tolerance and resilience of neural networks (FastPath workshop at ISPASS 2020).

Michael Connors (4th year undergraduate at Harvard University) August 2018 - April 2019

Improving resiliency of deep neural networks for denser eNVM storage (Senior thesis)

Sreela Kodali (3rd year undergraduate and summer research intern)

May 2017 - August 2017

Human Activity Recognition Using Wearables and a Low-power Deep Learning Accelerator (TechCon 2017)

SEMINAR AND INVITED TALKS

Sustainable Computing

Udit Gupta

Chapter of IEEE Society on Social Implications of Technology, Washington DC (October 2022)

DeepRecSys: A System for Optimizing End-To-End At-scale Neural Recommendation Inference

Udit Gupta

Specialization with Benchmarks for Emerging Applications co-located with MICRO 2022 (October 2022)

Sustainable Computer Systems: Modeling and Design

David Brooks and Udit Gupta

Microsoft (September 2022)

Understanding and Optimizing the Environmental Footprint of Computing

Carole-Jean Wu and Udit Gupta

Open Compute Project Global Summit (November 2021)

Designing Specialized Systems for Deep Learning-based Personalized Recommendation

Yale University, October 2021 (hosted by Professor Abhishek Bhattacharjee)

Designing Specialized Systems for Deep Learning-based Personalized Recommendation

Boston University, October 2021 (hosted by Professor Ajay Joshi)

Chasing Carbon: Going Beyond Efficiency to Understand the Elusive Environmental Footprint of Computing

Google Brain, August 2021 (hosted by Dr. Emma Wang)

Chasing Carbon: The Elusive Environmental Footprint of Computing

CLEAR workshop at ISCA 2021

Designing Specialized Systems for Deep Learning-based Personalized Recommendation

Cornell Computer Systems Lab, April 2021 (hosted by Professor Zhiru Zhang)

Designing Systems for Data Center Scale Recommendation

ARM, April 2021 (hosted by Dr. Paul Whatmough)

DeepRecSys: A System for Optimizing End-To-End At-scale Neural Recommendation Inference

PeRSonAl workshop at MLSys 2021

A Hands On Tutorial Using DeepRecSys to Optimize At-Scale Neural Recommendation Inference

Udit Gupta and Samuel Hsia

PeRSonAl workshop at ISCA 2020

At-scale Inference for Recommendation Systems

PeRSonAl workshop at ASPLOS 2020

MASR: Modular Accelerator for Sparse RNNs

Cognitive Architectures workshop at HPCA 2020