Udit Gupta

✉ ugupta@g.harvard.edu • ugupta.com

RESEARCH OVERVIEW

Spanning across computer architecture, systems, and machine learning, my research co-designs solutions across the computing stack to design more efficient, scalable, and environmentally responsible applications and systems.

Systems for Machine Learning: Developing specialized systems for deep neural networks, data center scale AI, and personalized recommendation engines

Computer Architecture: Designing application specific hardware at mobile and data center scale, building tools and infrastructure for hardware design space exploration, and using high level synthesis for accelerator implementations

Sustainable Computing: Quantifying the carbon footprint of computing across hardware life cycles, building tools for architectural carbon accounting, and investigating hardware/software methods for sustainable AI

My research has been published in top-tier systems and architecture conferences including ISCA, MICRO, HPCA, and ASPLOS. My work has also been recognized with an IEEE MICRO Top Picks honorable mention, as well as best paper award nominations at the PACT and DAC conferences.

EDUCATION

Harvard University, Ph.D. — Cambridge, MA
Computer Science — 2016-Present
Advisors: Professor David Brooks and Professor Gu-Yeon Wei

Harvard University, M.S. — Cambridge, MA
Computer Science — 2020

Cornell University, B.S. — Ithaca, NY
Electrical & Computer Engineering, Computer Science — 2012-2016
Advisor: Professor Zhiru Zhang

HONORS AND AWARDS

IEEE MICRO Top Picks Honorable Mention	2021
Best Paper Nominee at Parallel Architectures and Compilation Techniques (PACT)	2019
Best Paper Nominee at Design Automation Conference (DAC)	2018
Harvard Smith Family Fellowship	2017
National Science Foundation (NSF) GRFP Honorable Mention	2016
Richard A. Newton Young Fellows Scholarship at *DAC 2015*	2015
Cornell Eta Kappa Nu (HKN) - Electrical Engineering Honor Society	2013 - 2016
Cornell ECE Early Research Career Scholarship	2013

PROFESSIONAL INDUSTRY EXPERIENCE

Facebook AI Research (FAIR) — Menlo Park, CA
Visiting Research Scientist — October 2020-Present
Advisors: Dr. Carole-Jean Wu and Dr. Hsien-Hsin S. Lee
- Enabling system design for sustainable AI across model and hardware life cycles.

Facebook AI Research (FAIR) Menlo Park, CA
Research Intern January 2020 - October 2020
Advisors: Dr. Carole-Jean Wu and Dr. Hsien-Hsin S. Lee
- Designing data-center scale specialized accelerators for multi-stage recommendation pipelines.

Facebook AI Infrastructure Menlo Park, CA
Research Intern September 2018 - January 2020
Advisors: Dr. Carole-Jean Wu, Dr. Hsien-Hsin S. Lee, and Dr. Kim Hazelwood
- Characterizing the architectural implications of deep learning based personalized recommendation systems.
- Designing inference schedulers to optimize recommendation performance in datacenters with varying application-level requirements, model architecture, and hardware platforms.

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.

PEER-REVIEWED PUBLICATIONS

RecPipe: Co-designing Models & Hardware to Jointly Optimize Recommendation Quality & Performance
Udit Gupta, Samuel Hsia, Jeff Zhang, Mark Wilkening, 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 inference schedulers on commodity hardware and designing specialized accelerators.
Artifact Badges: Available, Functional, and Reproducible
Code available on GitHub

RecSSD: Near Data Processing for Solid State Drive Based Recommendation Inference
Mark Wilkening, **Udit Gupta**, Samuel Hsia, Caroline Trippel, Carole-Jean Wu, David Brooks, Gu-Yeon Wei
International 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.
Featured by: Harvard Gazette, Tech at Facebook, Bloomberg Green, The Guardian

Cross-Stack Workload Characterization of Deep Recommendation Systems
Samuel Hsia, **Udit Gupta**, Mark Wilkening, 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

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 2020 - 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 25mm^2 SoC with a 54.5$\times$ 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 syntehsis 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
Outlining the state and charting paths for future mentoring programs in computer architecture.

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
Quantifying 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

Most of computing's carbon emissions are coming from manufacturing and infrastructure
Carole-Jean Wu, **Udit Gupta**
Tech at Facebook, March 2021

Optimizing Infrastructure for Neural Recommendation At-scale
Carole-Jean Wu, **Udit Gupta**
Facebook AI, February 2020

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

Designing AI-Enabled Technology for Society
Udit Gupta, Lillian Pentecost
Harvard Science in the News (SITN), October 2018

Software Programmable FPGAs
Udit Gupta
Circuit Cellar (*Tech the Future* series), June 2017

PRESS

The Computer Chip Industry Has a Dirty Climate Secret
The Guardian, September 2021

The Chip Industry Has a Problem With Its Giant Carbon Footprint
Bloomberg, April 2021

Smaller, Faster, Greener
Harvard Gazette, March 2021

Facebook Open Sourced this Architecture for Personalized Neural Recommendation Systems
Medium (Data Series), May 2020

OPEN SOURCE TOOLS AND INFRASTRUCTURE

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

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

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

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
Helped build the initial MLPerf benchmark and guide recommendation inference benchmark design
Code available on MLPerf

PROFESSIONAL SERVICE AND COMMUNITY INVOLVEMENT

Computer Architecture Student Association (CASA)
Steering Committee Member 2020-Present

- Student group fostering an invlusive 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.

Personalized Recommendation Systems and Algorithms (PeRSonAl) workshop
Co-founder and Organizer

- Co-founded workshop and tutorial to design personalized recommendation engines across application, algorithms, and systems and hardware.
- Hosted PeRSonAl workshop at top-tier systems and machine learning venues (**ASPLOS 2020**, **ISCA 2020**, and **MLSys 2021**) with over 200 attendees across over 50 institutions worldwide, 12 invited talks, and 15 contributed papers.

Computing Landscape for Environmental Accountability and Responsibility (CLEAR) workshop
Co-founder and Organizer

- Co-founded workshop and tutorial to design environmentally sustainable hardware systems across systems, hardware, and circuits.
- Hosted CLEAR workshop at top-tier systems and machine learning venues (**ISCA 2021**) with over 50 attendees, 9 invited talks, and an industry-academic panel.

Journal of Opportunities, Unexpected Limits, Retrospectives, and Experiences (JOURNE) workshop
Co-Founder and Organizer

- Hosted workshop on research journeys, unexpected paths and limitations, retrospectives, and experiences at **MLSys 2021**.

Negative results, Opportunities, Perspectives, and Experiences (NOPE) workshop
Co-Organizer

- Hosted workshop on sharing negative outcomes, post-mortems, and experiences in research at **ASPLOS 2021** and **ASPLOS 2019**.

Conference Review Committees

- International Conference on Learning Representations (*ICLR*) 2019 reviewer — ICLR 2019

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
- CS 141: Computing Hardware — Spring 2019

Undergraduate Teaching Assistant — Ithaca, NY
Cornell University

- CS 3420/ECE 3140: Embedding Systems — Spring 2016
- ECE 2300: Introduction to Digital Logic and Computer Organization — Fall 2015
- ECE 2300: Introduction to Digital Logic and Computer Organization — Spring 2015
- ECE 2300: Introduction to Digital Logic and Computer Organization — Spring 2014

edX Course Development — Ithaca, NY
Cornell University

- Helped design and develop EdX MOOC course *"The Computing Inside Your Smartphone"* — Summer 2014

Education Outreach for Middle School and High School Students

- Taught 1.5 hour class on *"Sustainable computing"* to 25 high school students at *Rainstorm* in Summer 2021.
- Taught 1 hour class on *"Sustainable computing"* to 30 high school students at *MIT Splash!* in Spring 2021.
- Taught 3 hour hardware engineering course to 25 high- school students at *Cornell Splash!* in Spring 2015.
- Taught 3 hour hardware engineering course to 30 high- school students at *Cornell Splash!* in Fall 2014.

RESEARCH MENTORING

Samuel Hsia (1^{st} year PhD Student) — August 2019 - Present
Characterizing hardware performance of neural recommendation models using Intel TopDown (*IISWC 2020*)

Liu Ke (3^{rd} year PhD student) — May 2019 - Present
Designing near memory processing accelerators for data center scale personalized recommendation (*RecNMP, ISCA 2020*)

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

Javin Pombra (3^{rd} year undergraduate at Harvard University) — May 2020 - Present
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 (2^{nd} year undergraduate at Harvard University) — May 2020 - September 2020
Designing and building a training model zoo for neural recommendation.

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

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

Ted Pyne (3^{rd} 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 (4^{th} year undergraduate at Harvard University) August 2018 - April 2019
Improving resiliency of deep neural networks for denser eNVM storage (*Senior thesis*)

Sreela Kodali (3^{rd} 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

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