

# Udit Gupta

✉ [ugupta@g.harvard.edu](mailto:ugupta@g.harvard.edu) • 🌐 [ugupta.com](http://ugupta.com)

## Education

---

### Harvard University, Ph.D.

Computer Science

Cambridge, MA

2016-Present

Advisors: Professor David Brooks, Professor Gu-Yeon Wei

**Research Interests:** Computer architecture, sustainable computing, deep learning, personalized recommendation

### Harvard University, Masters of Science

Computer Science

Cambridge, MA

2020

GPA: 3.87

### Cornell University, Bachelor of Science

Electrical & Computer Engineering, Computer Science

Ithaca, NY

2012-2016

Advisor: Professor Zhiru Zhang

GPA: 4.00, Dean's List (All semesters), *summa cum laude*

## Research Experience

---

### Harvard University

Graduate Researcher

Cambridge, MA

2016-Present

- Detailing the environmental impact of computing across hardware life-cycles for mobile and data center scale systems.
- Accelerating DNN-based personalized recommendation with specialized schedulers, memory systems, and hardware.
- Developed benchmarks for DNN-based recommendation models based on in-depth architectural characterization.
- Designed specialized hardware to parallelize static and dynamic sparse execution in RNNs for on-chip speech recognition.
- Collaborated with graduate students and post-docs on 16nm tape-out with ARM A53 CPU and 4 coherent accelerators.

### Cornell University

Undergraduate Researcher

Ithaca, NY

2013-2016

- Developed benchmarks and optimizations for designing accelerators using high-level synthesis on FPGAs.

## Industry Experience

---

### Facebook, Inc.

Facebook AI Research (FAIR) Visiting Research Scientist

Menlo Park, CA

January 2021-Present

Facebook AI Research (FAIR) Intern

January 2020 - December 2020

AI Infrastructure Research Intern

September 2018 - January 2020

- Characterizing the architectural implications of deep learning based personalized recommendation systems.
- Designing inference schedulers and specialized hardware to optimize the performance of recommendation in datacenters under different run-time configurations such as models, server architecture, batching, and co-location.

### Algo-Logic Systems

Hardware Design and Verification Engineering Intern

Santa Clara, CA

Summer 2015

- Designed and implemented OpenCL interface to software kernels with existing IP on FPGAs for financial data parsers.

## 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 GRFP Honorable Mention	2016
Richard A. Newton Young Fellows Scholarship at DAC 2015	2015

## Open Source Initiatives

---

- DeepRecSys: A System for Optimizing End-To-End At-scale Neural Recommendation Inference ([GitHub](#))
- Ares: A framework for quantifying the resilience of deep neural networks ([GitHub](#))
- MLPerf: A Benchmark for Machine Learning from an Academic/Industry Cooperative ([MLPerf](#))

## Community Involvement and Professional Service

---

### Computer Architecture Student Association (CASA)

Steering Committee Member 2020-Present

- Steering committee member for a student-run group fostering an inviting computer architecture student community.
- Organizing a [summer 2021 reading group](#) focused on advancing and promoting diversity, equity, and inclusion in computer architecture.
- Gathered qualitative and quantitative feedback on various mentoring and networking initiatives in computer architecture, published at [Workshop on Computer Architecture Education \(WCAE\)](#) co-located with ISCA 2021.

### Workshop and Tutorial Organizing Activities

- Computing Landscapes for Environmental Accountability and Responsibility ([CLEAR](#)), Co-founder ISCA 2021
- Negative Outcomes Post-Mortems and Experiences ([NOPE](#)), Co-organizer ASPLOS 2021
- Journal of Retrospectives, Negative Results, Experiences ([JOURNE](#)), Co-founder MLSys 2021
- Personalized Recommendation Systems and Algorithms ([PeRSONAI](#)), Co-founder MLSys 2021
- Personalized Recommendation Systems and Algorithms ([PeRSONAI](#)), Co-founder ISCA 2020
- Personalized Recommendation Systems and Algorithms ([PeRSONAI](#)), Co-founder ASPLOS 2020
- Negative Outcomes Post-Mortems and Experiences ([NOPE](#)), Co-organizer ASPLOS 2019

### Professional Memberships

- Computer Architecture Student Association, Steering Committee Member 2020-Present
- Harvard SEAS Graduate Council, Member 2019-Present
- 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

Harvard University Cambridge, MA  
2 semesters

- CS 290: PhD Grad Cohort Research Seminar Fall 2020
- CS 141: Computing Hardware Spring 2019

### Research Mentor for Undergraduate Students

Harvard University Cambridge, MA

- Advised 3 summer undergraduate students on building recommendation training zoo.
- Advised undergraduate student on "*Quantifying the Impact of Data Encoding on DNN Fault Tolerance*" (*Fastpath workshop*).
- Advised undergraduate senior thesis on "*Improving Resiliency of Deep Neural Networks for Denser eNVM Storage*".
- Mentored 2 summer undergraduate students on "*Applications of Deep Neural Networks for Ultra Low Power IoT*" (ICCD 2017) .

### Education Outreach for Middle School and High School Students

Harvard University Cambridge, MA

- Taught a one-hour class on "*Sustainable computing*" to 30 high school students at [MIT Splash!](#) in Spring 2021.
- Taught three-hour introductory hardware engineering course, "*Computer's Don't Byte!*", to 25 middle-school and high-school students at [Cornell Splash!](#) in Fall 2014, Spring 2015.

## Undergraduate Teaching Assistant

Cornell University

- CS 3420/ECE 3140: Embedding Systems
- EdX MOOC: The Computing Inside Your Smartphone
- ECE 2300: Introduction to Digital Logic and Computer Organization

Ithaca, NY

4 semesters

Spring 2016

Summer 2014

Spring 2014, Fall 2015, Spring 2015

## Publications

---

*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)

*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)

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

*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)

*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).

*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).

*Architectural Implications of Facebook's DNN-based Personalized Recommendation*

**Udit Gupta**, Xiaodong Wang, Maxim Naumov, Carole-Jean Wu, Brandon Reagen, David Brooks, Bradford Cattel, 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)

★ *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).

★ *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 16nm 25mm<sup>2</sup> 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)

*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).

*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).

*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).

★ *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).

*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)

*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)

*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)

## Technical Articles

*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

*Designing AI-Enabled Technology for Society*

**Udit Gupta**, Lillian Pentecost

Harvard SITN, October 2018