

Udit Gupta

10 Adams Court – Plainsboro, NJ 08536

• ☎ (609) 529 7670 • ✉ ugupta@g.harvard.edu • ✉ ugupta.com

Education

Harvard University , Ph.D.	Cambridge, MA
Computer Science	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	Cambridge, MA
Computer Science	2020
GPA: 3.87	
Cornell University , Bachelor of Science	Ithaca, NY
Electrical & Computer Engineering, Computer Science	2012-2016
Advisor: Professor Zhiru Zhang	
GPA: 4.00, Dean's List (All semesters), <i>summa cum laude</i>	

Research Experience

Harvard University	Cambridge, MA
Graduate Researcher	2016-Present
◦ Detailing the environmental impact of computing at mobile and data center scale.	
◦ 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	Ithaca, NY
Undergraduate Researcher	2013-2016
◦ Developed benchmarks and optimizations for designing accelerators using high-level synthesis on FPGAs.	

Industry Experience

Facebook, Inc.	Menlo Park, CA
Facebook AI Research (FAIR) Visiting Research Scientist	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	Santa Clara, CA
Hardware Design and Verification Engineering Intern	Summer 2015
◦ Designed and implemented OpenCL interface to software kernels with existing IP on FPGAs for financial data parsers.	

Honors and Awards

Honorable Mention for IEEE MICRO Top Picks	2021
Best Paper Nominee at Parallel Architectures and Compilation Techniques (PACT)	2019
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
<https://github.com/harvard-acc/DeepRecSys>
- MLPerf: A Benchmark for Machine Learning from an Academic/Industry Cooperative.
<https://mlperf.org/>
- Ares: A framework for quantifying the resilience of deep neural networks.
<https://alugupta.github.io/ares/>

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 (PeRSonAl), Co-founder MLSys 2021
- Personalized Recommendation Systems and Algorithms (PeRSonAl), Co-founder ISCA 2020
- Personalized Recommendation Systems and Algorithms (PeRSonAl), Co-founder ASPLOS 2020
- Negative Outcomes Post-Mortems and Experiences (NOPE), Co-organizer ASPLOS 2019

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 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)

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 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)

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

Teaching and Leadership Experience

Graduate Teaching Fellow	Cambridge, MA
Harvard University	2 semesters
o CS 290: PhD Grad Cohort Research Seminar	Fall 2020
o CS 141: Computing Hardware	Spring 2019
 Undergraduate Research Mentor	Cambridge, MA
Harvard University	
o Advised 3 summer undergraduate students on building recommendation training zoo.	
o Advised undergraduate student on " <i>Quantifying the Impact of Data Encoding on DNN Fault Tolerance</i> " (<i>Fastpath workshop</i>).	
o Advised undergraduate senior thesis on " <i>Improving Resiliency of Deep Neural Networks for Denser eNVM Storage</i> ".	
o Mentored 2 summer undergraduate students on " <i>Applications of Deep Neural Networks for Ultra Low Power IoT</i> " (ICCD 2017).	
 Undergraduate Teaching Assistant	Ithaca, NY
Cornell University	4 semesters
o CS 3420/ECE 3140: Embedding Systems	Spring 2016
o EdX MOOC: The Computing Inside Your Smartphone	Summer 2014
o ECE 2300: Introduction to Digital Logic and Computer Organization	Spring 2014, Fall 2015, Spring 2015
 IEEE Student Chapter	Ithaca, NY
President and Corporate Director	2013-2016
o Recruited and led 28 undergraduate and graduate students to organize corporate, social, and outreach events.	
o Led 5 students to administer a <i>Cornell Splash!</i> class, " <i>Computers Don't Byte</i> ", to 24 high school students.	

Professional Memberships and Service

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
Harvard SITN Blog (Editor)	2018-2019