Udit Gupta

10 Adams Court – Plainsboro, NJ 08536
(609) 529 7670 • ugupta@g.harvard.edu • http:/www.ugupta.com

Education

Harvard University, Ph.D. — Cambridge, MA
Computer Science — 2016-Present
Advisors: Professor David Brooks, Professor Gu-Yeon Wei
GPA: 3.87
Research Interests: Computer architecture, machine learning, deep neural networks, hardware accelerators

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), *summa cum laude*

Research Experience

Harvard University — Cambridge, MA
Graduate Researcher — 2016-Present

- Designing specialized hardware for RNN based automatic speech recognition. Devised a low-cost sparse encoding technique to exploit sparsity in activations and weights for scalable parallelism and real-time, on-chip speech recogniton using state-of-the-art RNNs.
- Designed and integrated, using high-level synthesis, neural network accelerators with mobile SoCs. Collaborated with graduate students and post-docs, to tape out chip in 16nm technology with ARM A53 CPU coherent with 4 accelerators.
- Developed techniques to lossily compress neural networks by up to 500× using probabilistic data structures.
- Investigated reliability in deep learning, for various neural networks, with a large scale empirical study.

Cornell University — Ithaca, NY
Undergraduate Researcher — 2013-2016

- Evaluated hardware acceleration on FPGAs using high-level synthesis (C/C++ and OpenCL).
- Evaluated scheduling, mapping, and memory optimizations for high-level synthesis (co-authored 2 publications).
- Led team of 5 undergraduate students to build an application level benchmark suite for high-level synthesis.

Industry Experience

Facebook, Inc. — Menlo Park, CA
AI Infrastructure Research Intern — September 2018-Present

- Studying the architectural implicaitons of deep learning based personalized recommendation systems.
- Analyzing and characterizing the performance of at-scale recommendation models in datacenters under different run-time configurations such as server architecture, batching, and co-location.

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

- Built an OpenCL board support package to interface software kernels with existing IP on FPGAs.
- Devised software API for configuring OpenCL based FPGA financial data parsers.

Open Source Initiatives

- 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/

Publications

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
Under review.
https://arxiv.org/abs/1906.03109

MASR: A Modular Accelerator for Sparse RNNs
Udit Gupta, Brandon Reagen, Lillian Pentecost, Marco Donato, Thierry Tambe, Alexander Rush, Gu-Yeon Wei, David Brooks
To appear in 2019 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.
To appear in 2019 IEEE/ACM International Symposium on Microarchitecture (MICRO 2019).

A 16nm 25mm^2 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.
2019 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.
30^{th} 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.
35^{th} International Conference on Machine Learning (ICML 2018).
https://arxiv.org/abs/1711.04686

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.
55^{th} Design Automation Conference (DAC 2018). ***Best Paper Nominee***
https://alugupta.github.io/ares/

On-chip Deep Neural Network Storage with Multi-level eNVM.
Marco Donato, Brandon Reagen, Lillian Pentecost, **Udit Gupta**, David Brooks, Gu-Yeon Wei.
55^{th} 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.
26^{th} 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.
25^{th} 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.
23^{rd} ACM/SIGDA International Symposium on Field-Programmable Gate Arrays (FPGA 2015)

Teaching and Leadership Experience

Undergraduate Research Mentor
Cambridge, MA

Harvard University

- Advised undergraduate senior thesis on *"Improving Resiliency of Deep Neural Networks for Denser eNVM Storage"*.
- Mentored 2 summer undergraduate student projects in deep learning on low power IoT devices published at the International Conference on Computer Design (ICCD 2017), *"Applications of Deep Neural Networks for Ultra Low Power IoT"* .

Graduate Teaching Fellow
Cambridge, MA

Harvard University — 1 semester

- Designed lecture materials, assignments, exams, and final project for introductory computing hardware class (CS141).
- Conducted office hours and graded assignements and exams for class of 50 students.

Undergraduate Teaching Assistant
Ithaca, NY

Cornell University — 4 semesters

- Conducted office hours and proctored labs of 30 students for hardware and software classes (ECE 2300 and CS3140) .
- Collaborated in designing assignments and labs for an online (EdX) course, *"The Computing Inside Your Smart Phone"*.

IEEE Student Chapter
Ithaca, NY

President and Corporate Director — 2013-2016

- Recruited and led 28 undergraduate and graduate students to organize corporate, social, and outreach events.
- Led 5 students to administer a *Cornell Splash!* class, *"Computers Don't Byte"*, to 24 high school students.

Honors and Awards

Harvard Smith Family Fellowship	2017
National Science Foundation GRFP Honorable Mention	2016
Richard A. Newton Young Fellows Scholarship at *DAC 2015*	2015
Cornell ECE Early Research Career Scholarship	2013
Eta Kappa Nu - Electrical Engineering Honor Society	2013-2016

Skills

- Programming: Python, C/C++, Verilog, OpenCL, Java
- CAD Tools: Catapult High-Level Synthesis, Xilinx Vivado Design Suite, Altera Quartus
- Machine Learning Libraries: Keras, PyTorch, Caffe2