Aayush Ankit

☎ *(765) 409 3526*
✉ *aankit@purdue.edu*
Linkedin - Aayush Ankit
https://github.com/Aayush-Ankit

5th Year Computer Engineering PhD Student at Purdue University
Research focus: Hardware and software design for efficient machine learning.
Areas of Interest/Skills: Computer Architecture, GPU, Hardware Accelerators, Hardware-Software Codesign, VLSI Design, Deep Learning. Machine Learning.

Education

2015–2020 (expected) **Direct Ph.D.**, *Electrical and Computer Engineering, Purdue University*, West Lafayette, IN, *GPA – 3.94/4.00*.

2011–2015 **B.Tech.**, *Electronics Engineering*, Indian Institute of Technology (BHU), Varanasi, India. *GPA – 8.72/10.00*

Graduate Research

Jan 2016 – Present **Graduate Research Assistant**, *Purdue University*, West Lafayette, IN.
Advisor: **Prof. Kaushik Roy**
Funded by JUMP C-BRIC center, Hewlett Packard Labs (part-time).

In the past, I have developed end-to-end machine learning inference accelerators with hybrid memristor-digital technology - RESPARC (DAC'17), PUMA (ASPLOS'19). I have also designed transformation algorithms for efficient inference of sparse neural networks with memristors - TraNNsformer (ICCAD'17). **My research has resulted in the first open-source architectural simulator for ReRam-based accelerators (puma-simulator)**. Following are my current projects (work-in-progress):

- **ML Training using ReRam (collaboration with *Hewlett Packard Labs*)**.
  - Exploring approximate writes to circumvent the high write cost of ReRam crossbars to harness the benefits of in-memory processing for ML training.
  - Developed functional simulator in TensorFlow to model key technology characteristics for exploring training flow convergence.
  - **Accepted at CogArch, HPCA'20.**
- **Analog SRAM based in-memory GPGPU architecture for throughput computing applications.**
  - GPGPU architectures obtain state-of-the-art performance on throughput oriented workloads.
  - Exploring in-memory computing enabled by analog SRAM to improve the efficiency of GPGPU by reducing accesses to Register File, Caches, Shared Memory and resulting data movements.

Internships

May 2019 – Aug 2019 **GPU Architect Intern *(Advanced Computing Labs)***, *Samsung Electronics* , San Jose, CA, *Mentors: Wilson Fung, Anders Kugler*.

- Worked on texture unit bottleneck analysis and proposed an approximate texture filtering technique to improve the efficiency of existing and emerging programs.
- Explored operand reuse for improving vector register file bandwidth.
- Performed studies in the production C++ architecture simulator used by product teams to define next generation architectures.
- Work resulted in filing one US patent application.

Sep 2017 – Dec 2017 **CPU Design Intern *(Intel Architecture Cores Group)***, *Intel Corporation* , Hillsboro, OR, *Mentors: James Hadley, Christopher Palistrant*.

- Investigated PPA and proposed microarchitecture techniques to improve instruction decode and steering stages, their RTL design and formal verification.
- Synthesis and PnR of existing blocks to analyze for timing failures and study the impact of different architectural choices.

May 2017 – Aug 2017 **ML Accelerator Architect Intern *(Systems Architecture Lab)***, *Hewlett Packard Labs* , Palo Alto, CA, *Mentors: Dejan Milojicic, John Paul Strachan*.

- Designed PUMA: an ISA-programmable and general-purpose architecture built with NVM crossbars that can implement all varieties of ML applications (CNN, LSTM, MLP etc.).
- Developed a cycle-level simulator (performance, power and functionality) for the proposed architecture for design space exploration and benchmarking.
- Proposed compiler optimizations for improving performance.
- Work has been adopted for Advanced Development projects at HPE (DPE-PUMA).

Selected Publications

1. **Aayush Ankit**, Izzat El Hajj, Sai Rahul Chalamalasetti, Sapan Agarwal, Matthew Marinella, Martin Foltin, John Paul Strachan, Dejan Milojicic, Wen-mei Hwu, Kaushik Roy "PANTHER: A Programmable Architecture for Neural Network Training Harnessing Energy-efficient ReRAM", *CogArch Workshop, International Conference on High Performance Computer Architecture*, ***(CogArch, HPCA 2020)***.
2. **Aayush Ankit**, Izzat El Hajj, Sai Rahul Chalamalasetti, Geoffrey Ndu, Martin Foltin, R. Stanley Williams, Paolo Faraboschi, Wen-mei Hwu, John Paul Strachan, Kaushik Roy, Dejan Milojicic, "PUMA: A Programmable Ultra-efficient Memristor-based Accelerator for Machine Learning Inference", *ACM International Conference on Architectural Support for Programming Languages and Operating Systems*, ***(ASPLOS 2019)*** (Acceptance rate: 74/350=21%).
3. Joao Ambrosi, **Aayush Ankit**, Rodrigo Antunes, Sai Rahul Chalamalasetti, Soumitra Chatterjee, Izzat El Hajj, Guilherme Fachini, Paolo Faraboschi, Martin Foltin, Sitao Huang, Wen-mei Hwu, Dejan Milojicic, Kaushik Roy, John Paul Strachan, et. al. "Hardware-Software Co-Design for an Analog-Digital Accelerator for Machine Learning", *IEEE International Conference on Rebooting Computing*, ***(ICRC 2018)***.
4. **Aayush Ankit**, Abhronil Sengupta, Kaushik Roy. "TraNNsformer: Neural Network Transformation for Memristive Crossbar based Neuromorphic System Design", *IEEE/ACM International Conference on Computer Aided Design*, ***(ICCAD 2017)*** (Acceptance rate: 105/399=26%).
5. **Aayush Ankit**, Abhronil Sengupta, Priyadarshini Panda, Kaushik Roy. "RESPARC: A Reconfigurable and Energy-Efficient Architecture with Memristive Crossbars for Deep Spiking Neural Network", *ACM Design Automation Conference*, ***(DAC 2017)*** (Acceptance rate: 161/676=24%).

For detailed list please check my google scholar profile: Aayush-Google-Scholar

Technical skills

- **Programming Languages:** ***Fluent***: C++, Python, Verilog, Matlab, CUDA. ***Medium***: pthreads, MPI. ***Basic***: Bash, Perl.
- **System Simulation:** gem5, GPGPU-Sim
- **Deep Learning:** Tensorflow, Torch
- **EDA Tools:** Synopsys Design Compiler, Synopsys IC Compiler, Cadence Jasper Gold
- **FPGA Prototyping:** Modelsim, Quartus, SignalTap Logic Analyzer, Xilinx Virtex 5, Altera DE2, miniBee.
- **Transistor/Layout Design and Analysis:** Nanosim, HSPICE, Cadence Virtuoso

Graduate Coursework

- **Systems and Architecture**: Computer Design and Prototyping (Fall'15), Computer Architecture (Fall'16), Embedded Systems (Spring'16), GPGPU and Programmable Parallel Accelerators (Spring'18), Compilers and Translator Writing Systems (Fall'18), Parallel Computer Architecture (Spring'19)
- **VLSI and Circuit**: MOS VLSI Design (Fall'15), Solid State Devices (Spring'16), Advanced VLSI Design (Spring'17), Digital Systems Design Automation (Spring'17)
- **Machine Learning and Math**: Deep Learning (Fall'16), Random Variables and Probability (Spring'18), Linear Algebra and Applications (Fall'19)
- **Online courses**: Machine Learning (Stanford), Object Oriented Programming (IIT-B), Introduction to Computer Science and Programming Using Python (MIT)

Academic Projects

Spring'19 **MSI and DSI based directory cache coherence**, ***[ECE 666]***.
Skills: C++, gem5, Ruby memory model

- Implemented the MSI (Modify, Shared, Invalid) and DSI (Dynamic Self Invalidation) coherence protocols for directory-based shared memory systems.
- Analyzed the impact of DSI w.r.t MSI on coherence traffic from reduction in invalidation(s) and increase in self-invalidation(s).

Fall'18 **Compiler for Tiny Language to MICRO Assembly (toy ISA)**, ***[ECE 573]***.
Skills: C++, Flex, Bison

- Implemented the front-end and back-end of compiler infrastructure - scanner, parser, abstract syntax tree construction (variable scope, type-check, activation records), IR generation and assembly code generation.
- Worked on compiler optimizations (register allocation, loop optimizations) to improve performance.

Spring'18 **Energy-Efficient GPGPU Register File Design**, ***[ECE 695]***.
Skills: C++, GPGPU-Sim, CUDA

- Implemented and analyzed Register File Caching (RFC) for capturing temporal locality in the working set of registers accessed by the warps.
- Designed In-Memory Computing (IMC) based register file to reduce the number of accesses to Register File.
- Proposed architectural techniques to enable IMC based register file design: operand alignment, enforcing bank-conflicts within warp.

Spring'17 **Design Automation of DNN for Accuracy-Cost Co-optimization**, ***[ECE 595Z]***.
Skills: Caffe, Python

- Explored a branch-and-bound based search algorithm for hyper-parameter search to automate the design of CNNs for co-optimizing accuracy and computational cost.
- Script to automate the construction of NN prototype files were designed in Python and the ConvNets were trained with Caffe on a GPU server consisting of TitanX GPUs.

Fall'16 **Iterative Synapse Pruning based DNN Training for Energy-Efficiency**, ***[BME 595]***.
Skills: Torch (nn, cunn)

- Implemented iterative pruning of deep neural networks to dynamically learn the architecture of a ConvNet along-with weight training by creating "nn.Prune" modules in Torch.
- Analyzed the pruning efficiency (fraction of synapses which can be removed) and the training effort (increase in training time) across multiple benchmarks on MNIST and CIFAR-10 dataset.
- Benchmarks were run on Nvidia GTX 950M using the Torch CUNN package for training and testing purposes.

Fall'16 **Significance-driven Cache Compression for Improved Cache Performance**, ***[ECE 565]***.
Skills: C++, gem5

- Implemented the compression and decompression engine to enable Frequent Pattern Compression (FPC) based cache compression on L1/L2 cache.
- Studied the resulting tradeoffs between higher cache capacity higher and higher access latency towards IPC improvements.

Fall'15 **Multicore Processor Design**, ***[ECE 565]***.
Skills: System Verilog, Modelsim, Quartus, SignalTap II Logic Analyzer, Altera DE2 FPGA

- Designed a shared memory based dual core processor using MSI coherency protocol.
- Hardware support for atomic Read-Modify-Write operations was added for synchronization between the cores.
- The hardware was designed using System Verilog, prototyped on Altera's DE2 FPGA development board and tested extensively using assembly files.

Fall'15 **Voltage Over-scaling by Unbalanced Pipelining, Wallace Tree Multiplier**, ***[ECE 559]***.
Skills: Cadence Virtuoso, HSPICE, Nanosim, Cosmoscope

- Schematic design, layout and timing analysis of an 8-bit Multiplier with unbalanced pipeline using self-developed standard-cell libraries in Cadence Virtuoso.
- The pipeline uses critical path isolation technique for timing adaptiveness to achieve low-power variation-tolerant design.
- Power and performance of the design was measured using post layout simulation.

Professional Affiliations

Student Member Institue of Electrical and Electronics Engineers (IEEE), Semiconductor Research Corporation (SRC), Center for Brain-Inspired Computing Enabling Autonomous Intelligence (C-BRIC) - an SRC and DARPA sponsored JUMP center

Reviewer DAC (expert reviewer), ICCAD (external reviewer), IEEE journals (TVLSI, TCAD, TETC, TNNLS, TCAS-2, Access), ACM journals (JETCAS, JETC)

Academic Achievements

2018 IEEE SiPS 2018 Best Paper Award

2015 Awarded Edward Tiedemann scholarship and Teaching Assistantship for graduate studies at Purdue University

2014 Mitacs Globalink Research Fellowship - selected out of 4000 applicants across multiple countries to pursue a funded research project in University of Alberta, Canada for 3 months

2013 South Korea Summer Fellowship - selected out of 500 applicants from IIT-BHU to pursue a funded research project in Hanyang University, South Korea for 2 months

2012 Secured All India Rank 10 in SCRA, UPSC out of 190,000 candidates and appointed for Class-1 officer (highest entry-level designation), Government of India

2011 Secured All India Rank 2668 in IIT-JEE out of 470,000 candidates (top 0.5%)

2009 Selected for HBCSE-TIFR camp, International Junior Science Olympiad - among top 36 finalists

Residency Status

Nationality Indian

Visa F1 (since August 2015)