Shubham Negi

Contact Information
2550 Yeager Road, Apt # 13-6
West Lafayette IN, USA 47906

Electronic Information
snegi@purdue.edu (765) 714 8068
Google Scholar LinkedIn GitHub

I am a fifth-year Computer Engineering PhD student at Purdue University. My project experience includes a range of work in machine learning, focusing on hardware, software and algorithm development. I am particularly interested in hardware software co-design for machine learning workloads.

Education

- **Purdue University, West Lafayette** 2019 - Present
  Ph.D. Student, School of Electrical & Computer Engineering GPA: 4/4
- **Indian Institute of Technology Kharagpur** 2014 - Present
  B.Tech (Honors) + M.Tech (Dual Degree) in Electrical Engineering GPA: 8.94/10
  Specialization in Instrumentation and Signal Processing GPA: 9.34/10

Awards & Achievements

- **Ross Fellowship**: Awarded Ross fellowship for graduate studies at Purdue University. (2019)
- **S.N. Bose Scholar:** An extremely prestigious scholarship awarded to only 50 students from all over India, by SERB, Govt. of India, DST, Indo-US Science & Technology Forum (IUSSTF), & WINStep Forward, to fund research internship of Indian students in U.S. universities. (2018)
- **Inter IIT Tech. Meet:** Fifth position out of 100 models presented in Engineering Conclave event. (2018)
- **Department Rank 3** in Instrumentation and Signal Processing specialization.

Publications

- **Shubham Negi**, Indranil Chakraborty, Aayush Ankit, Kaushik Roy,“NAX: Neural Architecture and Memristive Xbar based Accelerator Co-design”, **DAC 2022**.
- **Shubham Negi**, Utkarsh Saxena, Deepika Sharma, Kaushik Roy,“HCiM: ADC-Less Hybrid Analog-Digital Compute in Memory Accelerator for Deep Learning Workloads”, **submitted to ISLPED,2024**.
- **Shubham Negi**, Deepika Sharma, Adarsh Kosta, Kaushik Roy,“Best of Both Worlds: Hybrid SNN-ANN Architecture for Event-based Optical Flow Estimation”, **submitted to IROS 2024**.
- Deepika Sharma, **Shubham Negi**, Trishit Dutta, Amogh Agrawal, Kaushik Roy, “A 2.5 - 5 TOPS/W Input Sparsity-aware Reconfigurable Digital Compute-in Memory Spiking Neural Network Accelerator Core for Event-based Perception”, **submitted to VLSI Symposium 2024**.

Projects

- ADCLess Hybrid Analog Digital Compute in Memory (CiM) Accelerator
  Skills - System Verilog, Cadence virtuoso, Performance Simulator
  - Analyzed bottlenecks in deploying DNN with binary/ternary quantized partial sums to CiM Accelerator.
  - Proposed an algorithm-hardware co-design approach to design an ADC-Less CiM accelerator.
  - Developed a Digital CiM macro to process scale factors with a novel in-memory subtraction technique. It also utilizes the inherent sparsity in ternary partial sums to further reduce energy consumption.
  - **Key Insight:** Achieved 8× and 4× improvement in energy compared to baseline architecture using 7-bit and 4-bit ADC with minimal drop in accuracy.
- Co-Designing Neural Network and In-Memory Computing Hardware Architecture
  Skills - Performance and Functional simulator, PyTorch, Profilers (nvprof, line profiler)
  - Analyzed the implications of crossbar size in IMC accelerators on hardware efficiency and accuracy.

  - Developed a hardware-aware Neural Architecture Search (NAS) framework in PyTorch to co-design neural network and IMC hardware architecture.
  - Improved the NAS algorithm's efficiency by enhancing parallelism in the functional simulator, employing tiling strategies to boost GPU utilization and reduce search times.
  - **Key insight**: Co-designing the neural network and hardware architecture gives better tradeoff between accuracy and energy efficiency.
- Efficient temporal information Processing using Hybrid SNN-ANN models
  Skills - PyTorch, Timeloop (analytical tool)
  - Compared the accuracy and energy efficiency of ANNs and SNNs deployed on ML accelerator.
  - Proposed Hybrid SNN-ANN networks to efficiently capture temporal information from inputs.
  - **Key insight**: SNNs outperform ANNs in accuracy when processing inputs with temporal information.

Internship

- **Systems Engineering Intern (Kilby Labs)**, Texas Instruments Dallas
  Mentors: Mahesh Mehendale, Hetul Sanghvi May 2022 - Aug 2022
  - Focused on developing a hardware-aware mixed precision quantization search framework for TinyML accelerators.
  - Conducted extensive explorations of diverse neural network topologies employing One-shot Neural Architecture Search techniques.
  - **Key Achievement**: Successfully integrated both the mixed precision quantization search and NAS frameworks into the company's PyTorch library, enhancing its functionality and versatility.

Technical Skills

- **Programming Language**: Python, C, C++, Verilog, Matlab, CUDA.
- **System Simulation**: gem5.
- **Deep Learning**: PyTorch.
- **EDA Tools** : Synopsys Design Compiler.
- **Transistor/Layout Design and Analysis** : Cadence Virtuoso, HSPICE.

Academic Coursework

- **Systems and Architecture**: Computer Architecture and Operating Systems (Fall'16), Computer Architecture (Fall'20), Compilers and Translator Writing Systems (Fall'21), Parallel Computer Architecture (Spring'22), Programmable Accelerator Architectures (Spring'23)
- **VLSI and Circuits**: Architectural Design of ICs, Analog VLSI circuits, Digital VLSI Circuits, VLSI CAD
- **Machine Learning and Math**: Machine Learning II (Fall'19), Deep Learning (Spring'20), Linear Algebra Applications (Spring'21)

Academic Projects

- Exploring different branch predictor topologies [Fall'2020]
  - Developed and integrated YAGS, Bi-Mode, and local branch predictors into the gem5 architecture.
  - Trained Convolutional Neural Networks for branch prediction using CBKP-4 speck2k6 benchmarks.
  - **Key insight**: Demonstrated that the Neural Network-based branch predictor significantly reduced the misprediction rate compared to YAGS, Bi-Mode, and local predictors.
- Implementing the FMESI Cache Coherence Protocol in gem5 [Spring'2022]
  - Implemented the FMESI protocol over the existing MESI protocol within the gem5 architecture.
  - This enabled efficient cache-to-cache transfer of on-chip cache blocks, optimizing data accessibility.
  - **Key insight**: The integration of the forward state into the MESI protocol resulted in a notable speedup on benchmarks from the PARSEC suite, demonstrating improved performance.
- Implementing AlexNet network in CUDA [Fall'2023]
  - Implemented convolution, pooling and GEMM operations in CUDA.
  - Optimized General Matrix Multiplication to reuse data at the register file level as well
  - **Key insight**: Data reuse in Convolutional Neural Network can provide tremendous speedup.