Navneet

PhD in Computer Science, Georgia Institute of Technology

+1-404-829-5190
krnavneet13@gmail.com
Navneet

Education

Georgia Institute of Tecnology **[2024 - present]**
PhD in Computer Science : **GPA : 4.0/4.0**
Indian Institute of Technology Bombay **[2020 - 2024]**
Bachelor of Technology with Honors in Electrical Engineering; **GPA: 9.1/10**

Publications and Conferences

- **Navneet**, Anubhav Bhatla, Biswabandan Panda **"The Maya Cache: A Storage-efficient and Secure Fully-associative Last-level Cache"**, *51st International Symposium on Computer Architecture '24*
- **HiDeS: Hierarchical Delta Sparsity for Diffusion LLM Acceleration through Algorithm–System Co-Design"**, *HPCA '26 (under review)*

Research Experience

Acceleration Diffusion Model Inference on NVIDIA GPU **[Jan'26 - Present]**
Guides:*Prof. Celine Lin,, Georgia Tech*
- Worked on **HiDes**, an algorithm-system co-design approach that utilizes hierarchical sparsity to accelerate block-diffusion model inference by doing selective recomputation, across 3 granularities: layer, token, and columns.
- Built custom CUDA kernels for attention and MLP for to accommodate selective recomputation.
- Built our solution on top of **Fast-dLLM v2** block diffusion model achieving a speedup of **2.33x** and **2.67x** on **A100** and **H100** GPUs respectively.

Modeling DNN operators for ASTRA-sim using machine learning **[May '25 - Aug '25]**
Guides:*Prof. Tushar Krishna, Georgia Tech*
- Working towards accelerating **ASTRA-sim**, an open source infrastructure to model complete hardware-software co-design stack of distributed machine learning training systems
- Profiling opeators like **GEMM, Collective** etc, for state-of-the art LLMs on Nvidia GPUs using **CUDA**
- Developing Machine learning models to replace **Scale-Sim**, a systolic Array simulator used by **Astra-sim**, to estimate the execution timings of an arbitrary **GEMM** operators.

Assessed security vulnerabilities of ARM desktop CPUs **[Aug '24 - April'25]**
Guides:*Prof. Daniel Genkin,, Georgia Tech*
- Worked towards mounting side-channel attacks via reverse engineering ARM-based **Snapdragaon** desktop CPUs
- Reverse engineered microarchitectural components like prefetchers, RSB using custom **kernel-modules** on **Snapdragon X Elite** processors to mount attacks like **RSB-based Spectre**.

Avatar Cache: Enabling On-Demand Security with Morphable Cache Architecture **[June '24 - Aug '25]**
Guides:*Prof. Moinuddin K. Qureshi,, Georgia Tech, Prof. Biswabandan Panda, IIT Bombay*
- Proposed a **morphable** cache design to provide security **on-demand** with state-of-the-art cache design complexity.
- **On-demand** security feature allows users to opt out of security and hence avoid performance and power overheads.
- Idea is motivated by **Mirage** cache to provide extra ways for security but with high associativity to reduce overheads.
- Implemented the design on the **ChampSim** multi-core simulator, demonstrating performance overhead within 0.2% with 2% power overhead and 0.9% storage overhead when operating in the secure mode.

Storage-efficient and Secure Fully-associative Last-level Cache **[Jan '23 - Mar '24]**
Guide: *Prof. Biswabandan Panda, Computer Science, IIT Bombay*
- Proposed a secure and storage-efficient **shared LLC** to mitigate eviction-based **side-channel attacks**
- The design incorporates a decoupled tag and data store in the LLC, utilizing pointer-based indirection for 0% storage overhead compared to the non-secure baseline LLC, with minimal performance impact
- Idea is motivated by **Reuse** cache and **Mirage** cache which enabled us to reduce the data store size by 50% and keep the security similar to **Mirage** without hurting the performance too much

Internships

Advanced Micro Devices (AMD) | Research Associate - PhD **[Aug '25 - Dec' 25]**
- Implemented execution of GPU kernels on concurrent streams with inter-kernel data dependency to analyze performance gains from relaxing global synchronization constraints.
- Leveraged memory access patterns and inter-kernel thread block dependencies for **Lllama-3.1** in **vLLM** with a complete **Triton** backend to identify dispatch opportunities for concurrent kernel execution.
- Developed fine-grained dispatch strategies using global semaphores to overlap data-dependent kernel execution while preserving program correctness, targeting improved throughput and reduced dispatch overheads.

Texas Instruments | Digital Systems Intern **[May '23 - Jul '23]**
- Performed extensive analysis of interrupt signals of **RADAR SOC** built for automobile application

- Developed verilog logic for interrupt signals to detect **pulse-triggered** or **level-triggered** behaviour
- Designed a flow using **SystemVerilog Assertion** to detect any anomaly in the early stages of design and testing
- Integrated an automated process of monitoring and generating the status for all interrupt signals using **Python** and **SystemVerilog** in the main development pipeline

NanoSniff Technologies | **Embedded Systems Intern** **[Dec '21 - Jan '22]**
- Developed a solar-powered wireless data acquisition system using two **ESP32 wroom** module
- Designed charging circuit for Li-ion batteries using a dynamic power path management system with solar panels for uninterrupted power supply for a remote ESP32 node
- Created a **Python GUI** to collect data from the remote esp32 node and maintain a database using **PostgreSQL**

Technical Projects

Design of a trace-based Out-Of-Order CPU simulator **[Aug '24 - Dec '24]**
Guides:*Prof. Moinuddin K. Qureshi,, Georgia Tech*
- Designed a light-weight trace-based cycle-accurate 4-way wide **Out-of-Order CPU** simulator using C++.
- Modeled **multi-level cache hierarchy** and various cache replacement policies like **LRU, LFU** etc.
- Implemented cache partition techniques, including static schemes like way-partitioning and dynamic partitioning techniques like **Utility-based cache partitioning**

Design and Simulation of an Out-of-Order Superscalar Processor **[Sep '22 - Jan '23]**
Guide: Prof. Virendra Singh, Computer Science and Engineering, IIT Bombay
- Coordinated a four-member team to design a **2-wide superscalar Out-of-Order Processor**
- Designed logic for 2-bit branch predictor, Fetch, Decode, Reservation station, ROB along with register renaming
- Implemented the complete design of the processors and **tested** it on **VHDL** using **Quartus Prime**

Hazardous Gas detection using Nanosaur with VSLAM **[Jan '23 - Apr '23]**
Guide: Prof. Siddarth Tallur, Electrical Engineering, IIT Bombay
- Designed an add-on electronic circuit to detect gases like NH3, CO2, CO, CH4, etc. for **Nvidia Jetson Nano**
- Implemented a VSLAM technique; **ORBSLAM-2** with a stereo camera on jetson nano
- Integrated the sensor system and the **ORBSLAM-2** with a **Jetson Nano** based robot, Nanosaur, using **ROS2**
- Established remote control with the robot over wifi and set up a local server-based dashboard to collect gas intensity levels and show the 3D map of the environment

Fault Tolerant Core using FXA Architecture **[Aug '23 - Dec '23]**
Guide: Prof. Virendra Singh, Computer Science and Engineering, IIT Bombay
- Modelled **REMO**, a fault-tolerant core in **GEM5** using an in-order verifier for each instruction
- Modified the existing flow to model **FXA** architecture along with REMO fault tolerant feature
- Performed extensive testing and evaluated impact on performance using **SPEC06** benchmark suite

Image Deblurring using Image Deconvolution | **Course Project** **[Feb '23 - Apr '23]**
Guide: Prof. Amit Sethi, Electrical Engineering, IIT Bombay
- Leveraged **Deep CNN** to restore blurred images, effectively reversing the effects of motion blur
- Utilized **transfer learning** to improve model performance, remove noise, minimize unwanted artifacts and accelerate convergence by fine-tuning the **pre-trained CNN** on a dataset of 12240 image patches
- Achieved a significant reduction in unwanted artifacts while improving the image deblurring accuracy up to **92%** using a deep learning model built using **tensorflow** in python

Academic Achievements & Honours

- Awarded **Undergraduate Research Award** at IIT Bombay for my research on **Secure Last Level Cache** [2023]
- Received **Best Project Award** for building **Hazardous Gas Detection bot with VSLAM** [2023]
- Conferred with **KVPY (Young Scientist Incentive Plan)** fellowship with an All-India Rank **191** out of nearly 50,000 students in **SA-stream**, organised by **Indian Institute of Science, Bangalore** [2019]
- Achieved All-India Rank **649** out of 1,170,000 participants in **JEE (Main) 2020** [2020]
- Awarded **Certificate of Merit** in **NSEA** and **NSEJS** (National Standard Examination in Astronomy and Junior Science) for having been among **Top 1%** students among **53900+** candidates, organized by **IAPT** [2019, 2018]

Technical Skills

Languages	Python, C/C++, VHDL, Embedded C, Verilog, Assembly
Softwares & tools	vLLM, SgLang, Docker, GP-GPU Sim, GEM5, Tensorflow , Champsim, Quartus Prime, Git, PyTorch, Triton, CUDA

Key Courses

Architecture and Security	Hardware-Software co-design for machine learning systems, High performance computer architecture, Advanced Topics in Computer Architecture, Advanced Computer Architecture, Microprocessors , Principles of Data and System Security, Embedded Systems
Digital Design & VLSI	VLSI Design, Algoritmic Design of Digital Systems, Digital Systems
Machine Learning	Image Processing, Programming for Data Science, Intro to Machine Learning