SANJALI YADAV

Potomac, MD | sanjali7@umd.edu | sanjali.info | Linkedin | Github |

RESEARCH INTERESTS

AI Hardware Efficiency • Machine Learning for Systems • Sustainable Computing • Large Language Models

EDUCATION

University of Maryland, College Park

Ph.D. in Computer Science (GPA: 4.00/4.00) **Jan. 2024 – Dec. 2027**

M.S. in Computer Science (GPA: 4.00/4.00) **Jan. 2023 – Dec. 2023**

B.S. in Computer Science (GPA: 3.94/4.00) **Aug. 2019 – Dec. 2022**

PUBLICATIONS

Sanjali Yadav, Amirmahdi Namjoo, and Bahar Asgari. **Misam: Machine Learning-Assisted Dataflow Selection in Accelerators for Sparse Matrix Multiplication**. *MICRO 2025.*

Sanjali Yadav and Bahar Asgari. **Boötes: Boosting the Efficiency of Sparse Accelerators Using Spectral Clustering.** *MICRO 2025.*

Sanjali Yadav and Bahar Asgari. **DynaFlow: An ML Framework for Dynamic Dataflow Selection in SpGEMM Accelerators**. *IEEE Computer Architecture Letters Journal 2025.*

Amiramahdi Namjoo, Sanjali Yadav, Helia Hosseini, and Bahar Asgari. **Situla: Studying the Interplay of Sparse Formats and CPU/GPU Libraries.** ISPASS 2026.

RESEARCH EXPERIENCE

Graduate Research Assistant | (Advisor: Dr. Bahar Asgari) **Jan. 2024 - Present**

Computer Architecture and Systems Lab (CASL) **University of Maryland, College Park**

- **ML-Driven Optimization for Next-Gen AI** (*Myla*): Spearheading research on **Hardware-Software Co-design** for Generative AI workloads. My work utilizes machine learning techniques to optimize the execution of **Large Language Models (LLMs)** and **Mixture-of-Experts (MoE)** on resource-constrained GPUs.
- **Predictive Modeling** (*Misam and Dynaflow*): Engineered an ML-based runtime framework to predict optimal hardware configurations for dynamic inputs. This ML-driven approach solved the static architecture bottleneck, maximizing throughput and energy efficiency for irregular scientific and AI workloads.
- **Unsupervised Learning for Efficiency** (*Boötes*): Designed a data-efficiency pipeline using **Spectral Clustering** to reorder sparse matrices. By optimizing data locality, this system reduced off-chip memory traffic by **2.3x** and achieved **20x energy savings**.

Undergraduate Research Assistant | (Advisor: Ashok Agrawala) **Dec. 2019 – Dec. 2022**

Maryland Information and Network Dynamics Lab (MIND) **University of Maryland, College Park**

- **Big Data Analytics:** Architected a data ingestion pipeline to process and visualize high-frequency time-series data from **10,000+ IoT sensors** across campus buildings for energy propagation analysis, translating raw sensor logs into actionable sustainability insights.
- **Data Visualization:** Built an Augmented Reality (AR) interface (Unity/ArcGIS) to map real-time sensor streams to physical locations, enabling efficient large-scale facility management.

INDUSTRY EXPERIENCE

Capital One **Mclean, VA**

Software Engineering Intern *June 2024 – Aug. 2024*

- Engineered a **full-stack application** (Python Flask/Angular) to automate internal data registration workflows.
- Led UI/UX design and conducted end-to-end integration testing, iterating on user feedback to significantly reduce latency for business users.

Capital One **Mclean, VA**

Software Engineering Intern *June 2023 – Aug. 2023*

- Designed and built a scalable serverless data pipeline (**AWS SNS, Glue, DynamoDB, Lambda**) processing **1.5M email messages/year** for credit card customers.
- Integrated a predictive **ML model** for flight delay forecasting, enabling real-time market disruption assistance for customers.

Amazon **Seattle, WA**

Software Development Intern *May 2022 – Aug. 2022*

- Built the Cashback Abuse Prevention System (CAPS) in **Java and JavaScript** to automate the fraud detection of **20M users/year**.
- Deployed end-to-end production services using **AWS Lambda, AppConfig, and Elasticsearch**, protecting promotion campaigns from fraudulent accounts.

Amazon **Seattle, WA**

Software Development Intern *Jun. 2021 – Aug. 2021*

- Analyzed operational excellence practices and developed two new features for an internal tool using **Java and JavaScript**.
- Captured key operational metrics for **400+ teams** across Amazon, saving engineers **3 hours per week** by automating manual tracking processes.

AWARDS & HONORS

Finalist, Qualcomm Innovation Fellowship, North America, 2026

Machine Learning and Systems Rising Star, 2026

Certificate of Outstanding Achievement to Computer Science Graduate Program, 2026

First Place, ACM Student Research Competition (SRC) at **MICRO 2024** – *For Misam*

Capital One Bank Dean's Scholarship Fund in Computer Science

State of Maryland Governor's Award for Academic Excellence

ACADEMIC SERVICE

Artifact Evaluation Committee: MICRO 2025, ISCA 2025, ISCA 2026

Journal Reviewer: IEEE Micro (2025), IEEE Transactions on Big Data (2025), IEEE Transactions on Computer(2026)

Volunteer: SPICE Workshop at MICRO 2025

TECHNICAL SKILLS

Languages: Python, C++, Java, SQL, JavaScript, HLS, CUDA

AI & Data: PyTorch, Scikit-learn, Spark ML, LLM Inference

Cloud & Tools: AWS (Lambda, DynamoDB, Glue), Elasticsearch, FPGA, Git, Docker