Abhinav Nandwani

nandwani2@wisc.edu | linkedin.com/in/abhinavnandwani | github.com/abhinavnandwani | abhinavnandwani.com

EDUCATION

University of Wisconsin–Madison, College of Engineering Expected May 2026

Bachelor of Science in Computer Engineering (Machine Learning option) and Computer Science
GPA: 3.9/4.0

Relevant Coursework: Digital Design and Synthesis, Computer Architecture, Advanced Algorithms, Quantum System Architecture, Compilers, Design of Operating Systems

PROJECTS

YAPP Router UVM Testbench

Developed a full UVM testbench for a YAPP router with randomized packet generation, protocol integration, TLM scoreboarding, and register-level verification using Cadence Xcelium.

ARM-Based AI Chip

Deployed an 50M-parameter AI accelerator leveraging Meta's Llama 2 model on an AMD Xilinx Zybo Z7-20 FPGA using a custom 32×32 INT8 systolic array in SystemVerilog, achieving up to 250 MHz for low-power AI acceleration.

Pipelined RISC CPU

Co-designed a 5-stage MIPS CPU with register forwarding, dynamic branch prediction, and two-way set-associative caches; verified post-synthesis using Synopsys DC and static timing analysis.

Autonomous Inertial Robot

Led RTL design of an autonomous robot with inertial navigation using a NEMO gyro, UART/SPI/BLE interfaces, PID/PWM control, and 300 MHz low-power synthesis on a Cyclone IV FPGA

EXPERIENCE

Computer Architecture Researcher – Fault Resilience in Computer Vision Transformers

Prof. Pramesh Ramanathan, University of Wisconsin–Madison January 2025 – Present

- Investigated the increasing vulnerability of large transformer-based vision models to soft errors on GPUs, focusing on unprotected components in NVIDIA A100, and H100 architectures.
- Tasked with evaluating model reliability under hardware faults, designed experiments to assess the impact of instruction-level soft errors during real-world inference.
- Executed targeted hardware level fault injection campaigns, simulating bit-flips in SIMT pipelines, tensor core execution units, and warp schedulers.
- Discovered model-specific accuracy degradation in ViT, BEiT, and SegFormer across ImageNet, COCO, and Cityscapes; linked failure patterns to fault-prone microarchitectural structures, optimizing design strategies.

Computational Genomics Researcher – GPU Inference on single cell RNA AI workloads

Daifeng Wang Lab, University of Wisconsin–Madison February 2025 – Present

- Addressed the computational challenges posed by exploding single-cell RNA datasets by exploring scalable ML-based analysis methods.
- Applied LLM-scale transformer models to tokenized gene expression matrices for efficient, high-resolution single-cell clustering, classification and various other downstream tasks.
- Built optimized inference pipelines on NVIDIA A100 and H100 GPUs for 3M–100M parameter models, improving memory, batching, and device communication.
- Outperformed pretrained LLM baselines on clustering metrics (ARI, NMI, silhouette), contributing to a follow-up study with co-authors of a Nature-published Alzheimer's paper.

SKILLS

Languages: SystemVerilog, Verilog, VHDL, Python, C/C++, Java, Bash
Tools: Synopsys Design Compiler, Cadence Xcelium, Vivado, Quartus, ModelSim, Git, Linux
Hardware Expertise: RTL design, FPGA synthesis, GPU microarchitecture, memory hierarchy, pipeline profiling