Nikhil Rout

Hyderabad, India   nikhilrout97@gmail.com   +91 6303583185

nikhil-rout   NikhilRout   NikhilRout.github.io

Education

Vellore Institute of Technology, Chennai Campus *2022 – 2026*
Bachelor of Technology, Electronics and Communication Engineering *CGPA: 9.11/10*

Research Experience

Research Intern *Remote*
Vortex GPGPU, ORCAS Lab, UCLA *Mar 2025 – Present*

- Developed a Tensor Core Unit (TCU) extension for the open-source RISC-V based Vortex GPGPU, to accelerate WMMA operations for deep learning workloads
- Architected and implemented a configurable high-performance mixed-precision Fused Dot Product (FEDP) unit as part of the TCU in SystemVerilog
- Extended custom float conversions, SGEMM test, and cycle-level simulator implementation for FP8/BF8 support in C++
- Introduced a novel fused pipeline strategy for supporting Integer arithmetic within the floating-point datapath, maximizing component reuse while incurring minimal overhead
- Performed microarchitecture design space explorations, evaluating radix-4 booth recoding, 3:2 vs 4:2 compressor trees, and format-wise Dedicated vs Shared multiplier schemes
- Designed optimized Maximum Exponent and Modulo-4 operand grouping Carry-Save Adder modules to improve performance scaling at higher thread counts
- Reduced dynamic power consumption through clock/enable gating across FEDP pipeline stages during operation in 2:4 structured sparsity mode
- Contributed to TCU specific CI/CD and Verilator-based unittesting verification framework
- Achieved 4-cycle latency at 306.6 MHz on the Xilinx U55C FPGA, delivering ∼3.7× throughput over the open-source baseline (Berkeley HardFloat) at less than 60% the area cost

Summer Research Intern *Remote*
Centre for Nanoelectronics and VLSI Design, VIT Chennai *May 2024 – July 2024*

- Implemented pipelined radix-2 FFT modules using both DIT and DIF approaches
- Developed Number Theoretic Transform (NTT) modules for efficient polynomial multiplication in lattice-based post-quantum cryptography schemes on FPGAs

Research Publications

N. Rout, B. Tine, ”A Configurable Mixed-Precision Fused Dot Product Unit for GPGPU Tensor Computation,” *Vortex Workshop and Tutorials,* ***MICRO 2025***.

N. Rout, JJJ. Nesam, ”Optimizing RGB to Grayscale, Gaussian Blur and Sobel-Filter operations on FPGAs for reduced dynamic power consumption,” ***AIIoT 2024***.

Projects

Carry-Save FAN Microarchitecture DSE (SIGMA DNN Accelerator)

- Minimized Forward-Adder-Network (FAN) critical path by extending a Carry-Save-Adder tree structure for multi-vector reduction in sparse and irregular GEMM workloads
- Analyzed post-synthesis PPA design tradeoffs across multiple configurations

Flash-Attention CUDA MegaKernel

- Implemented a Flash-Attention Inference CUDA Megakernel with baseline and Tensor Core variants, leveraging shared memory tiling and online softmax to reduce HBM bottleneck
- Profiled kernel performance on NVIDIA T4 GPU, demonstrating speedup from operator fusion and memory hierarchy optimization

Skills and Competencies

Languages: Verilog/SystemVerilog, Chisel HDL, C/C++, CUDA, Python/PyTorch

Tools: Quartus Prime, Xilinx Vivado/Vitis, Synopsys VCS/DC, TCL, CMake, Git, LaTeX

Relevant Coursework: Digital Systems Design, VLSI System Design, Computer Architecture, FPGA-based System Design, Heterogeneous Computing Systems, Deep Learning

References

Dr. Blaise-Pascal Tine — **UCLA**
ORCAS Lab, Assistant Professor
✉ blaisetine@cs.ucla.edu

Dr. Jean Jenifer Nesam J — **VIT Chennai**
Assistant Professor
✉ jeanjenifernesam.j@vit.ac.in

Dr. Umadevi S — **VIT Chennai**
CNVD, Associate Professor
✉ umadevi.s@vit.ac.in