Soumil Paranjpay

Cupertino, CA | +1 (619) 953-7058
soumil07.com/linkedin | soumil.paranjpay@gmail.com

Education

University of California, San Diego Sep. 2024 – Dec. 2025

Master of Science in Electrical & Computer Engineering – CGPA: **3.78/4.00**

Vishwakarma Institute of Technology, Pune Aug. 2020 – Jun. 2024

B. Tech in Electronics & Telecommunication Engineering – CPI: **8.89/10.00** (Class Rank: 5/307)

Work Experience

Apple – GPU Architecture Modelling Engineer Cupertino, CA | Jan. 2026 – Present

- Functional and performance modelling of Apple GPUs

JPMorganChase – Software Engineering Intern Mumbai, India | Jan. 2024 – Jun. 2024

- Spearheaded the development of an **automation tool** for .NET Framework to Core migration, achieving a **95% reduction in migration times** through advanced code analysis.
- Collaborated with cross-functional teams to ensure seamless integration and adoption of the tool, significantly enhancing project efficiency.

Projects

Out-of-Order RISC-V Processor (C++) Jan. 2025 – Mar. 2025

- Architected and modelled a scalar Out-of-Order (OoO) processor incorporating **speculative execution** (GShare predictor), **dynamic instruction scheduling** via an instruction queue, and precise state management with **in-order retirement**.
- Developed a configurable C++ **performance simulator** to quantify the CPI impact of varying microarchitectural parameters (e.g., instruction queue depth, predictor table size) across different SPEC-CPU workloads.

Reconfigurable Systolic Array AI Accelerator (Verilog, Python) Sep. 2025 – Dec. 2025

- RTL Design, prototyping, and verification of a 16x16 systolic array AI accelerator, with reconfigurable SIMD and output-stationary modes for maximum flexibility.
- Trained quantized **VGGNet** and validated modified convolution layer to 16x16 accelerator tile.
- Mapped to Altera Cyclone FPGA and optimized for power and throughput with HW/SW codesign

Low-Power Dual Core Machine Learning Accelerator (Verilog) Jan. 2025 – Mar. 2025

- Designed and optimized RTL for a 16x16 **systolic array** for attention calculation.
- Implemented multi-VT place-and-route, **clock gating, power gating** to reduce power and improve PPA metrics by **45%.**
- Optimized RTL for sparse vector multiplication and implemented dual-core communication using async **4-way handshake** protocol.

Skills

Languages and Tools: C, C++, Python, Verilog, SystemVerilog, TCL, Gem5, Verilator