Paul Iacobucci

585-749-7932 | pmi22@cornell.edu | pauliacobucci.com | linkedin.com/in/pauliacobucci777 | github.com/pauliano22

Education

Cornell University Expected May 2027

Bachelor's in Computer Science; Minors in Artificial Intelligence, Applied Economics (Dyson) GPA: 3.64

Relevant Coursework: Machine Learning, Natural Language Processing, AI Reasoning, Systems Programming, Computer Networks, Embedded Systems, Data Structures & Algorithms, Functional Programming

Experience

Cornell Zhang Research Group Ithaca, NY

Machine Learning Research Assistant Jan. 2026 – Present

- Profiled frontier AI models on 8x H100 GPU clusters for an industry client designing custom AI silicon.
- Diagnosed expert-parallel bottlenecks via SGLang and Nsight, revealing 33% disparity and 117μs waits.
- Extended LLMServingSim with a per-layer Roofline model to classify compute- and memory-bound layers.
- Proved EP penalties scale strictly with intermediate dimension sizes via TP/EP scaling sweeps.

L3Harris Technologies Rochester, NY

Machine Learning Engineer Intern May 2026 – Present

- Built a real-time C++ text-to-speech AI system running fully on-device on an embedded military radio.
- Quantized a 15M-param VITS model Float32→INT8 via ARM NEON SIMD, cutting storage 70% to 17.79MB.
- Tuned ONNX Runtime memory arenas and stripped eSpeak NG to under 5MB, keeping active RAM below 45MB.
- Architected Yocto/BitBake recipes linking open-source dependencies with hardware vendor BSP layers.

Software Engineering Intern Dec. 2025 – Jan. 2026

- Engineered a Java streaming plugin for Android defense hardware, cutting live ATAK video feed latency by 20%.
- Offloaded real-time signal parsing from CPU to FPGA fabric, reducing system-wide overhead by 40%.

Software Engineering Intern May 2025 – Aug. 2025

- Developed a custom mapping and video streaming application for Linux-based military hardware.
- Extended a Go (Fiber) microservice sustaining 1,000+ concurrent streams and WebSocket chats using Goroutines.
- Achieved sub-1s video latency using MediaMTX to transmux WebRTC/RTSP/HLS for TCP/UDP connections.

Projects

FPGA-Accelerated MoE Trading Engine ↗ Apr. 2026

- Built an FPGA hardware trading engine generating trade signals from live stock exchange data in 444ns.
- Engineered a Vitis HLS pipeline executing a Sparse MoE router on live ITCH 5.0 market data feeds.
- Sustained 83.3M messages/second (50x over optimized C++) via O(1) parallel Limit Order Book lookups.
- Trimmed FPGA footprint below 5% utilization via 16-bit fixed-point arithmetic and DSP48 multiplications.

Triton GPU Performance Kernels (NVIDIA H100) ↗ Apr. 2026

- Wrote custom GPU kernels accelerating core AI model operations beyond native PyTorch performance.
- Fused FP8 quantization and LayerNorm on H100 SXM, hitting 3,906.1 GB/s bandwidth (45.7% speedup).
- Built a fused FlashAttention kernel using SRAM online softmax, scaling context to 16,384 tokens.

Mini-TensorRT: Deep Learning Graph Compiler ↗ Mar. 2026

- Built a from-scratch, zero-dependency C++ compiler that optimizes trained AI models for faster inference.
- Fused Conv-ReLU kernels to cut DRAM round-trips, measuring 11.2% speedup on cache-exceeding inputs.
- Matched ONNX Runtime logits exactly with handwritten multi-channel NCHW kernels on a trained CNN.

Leadership

Cornell Portfolio Management Club Ithaca, NY

Co-Founder & VP of Quantitative Research Sep. 2024 – Feb. 2025

- Co-founded a Cornell-recognized club bridging academic financial theory with practical portfolio management.
- Built the club website and a new-member curriculum of C++, Python, and quant interview prep resources.

Technical Skills

Languages: C++, Python, C, SystemVerilog, Go, Java, SQL, Bash

Hardware & Compute: CUDA, Triton, Vitis HLS, FPGA, PyTorch, TensorRT, ONNX Runtime, NCCL

Tools & Systems: Linux, Docker, Yocto/BitBake, Nsight, FastAPI, GitHub Actions, AWS

Interests: Poker, Chess, Powerlifting (315 lb bench), Golf