

Zephan Sanghani

zrs29@cornell.edu

(650) 521-3624

zephans.com

RESEARCH INTERESTS

My research interests are in energy-efficient computing and ML systems. I am interested in co-designing algorithms, architectures, and programming models to build efficient processors and accelerators, and in applying this approach to ML systems to make training and inference faster and cheaper at scale. My prior work has explored GPU energy efficiency for deep learning, manycore architectures, and FPGA acceleration.

EDUCATION

Cornell University, College of Engineering

Ithaca, NY

B.S. in Computer Science, Minor in ECE

Aug 2023 – May 2027

- GPA: 4.22/4.30
- Graduate coursework: High-Level Digital Design Automation, Memory Technologies and Memory-Centric Computing, Datacenter Architecture, Algorithmic Game Theory
- Undergraduate coursework: Analog IC Design, Digital VLSI Design, Computer Architecture, AI Hardware, Operating Systems, Analysis of Algorithms, Deep Learning

PUBLICATIONS

- [1] Po Jui Shih, **Zephan Sanghani**, Andrea Guarracino, Hasindu Gamaarachchi, and Christopher Batten. “Panomap: Unbiased Nanopore Signal Mapping with Pangenome Variation Graphs.” *bioRxiv* preprint, 2026. (*Under review.*)
doi:10.64898/2026.07.10.737796

RESEARCH EXPERIENCE

Batten Research Group, Cornell University

Ithaca, NY

Undergraduate Researcher, advised by Prof. Christopher Batten

Feb 2025 – Present

- Accelerated genomic dynamic programming on HammerBlade, a 1024-core RISC-V manycore, reaching 1.7× and 4.5× the area efficiency of CPUs and GPUs, respectively; now developing abstractions that simplify manycore programming and data movement.
- Explored graph-simplification approaches and ran experiments for Panomap [1], including adapting ODGI’s path-guided SGD layout for real-time chain deduplication; accelerated sequence and subsequence dynamic time warping for nanopore signals on FPGAs.

TEACHING EXPERIENCE

Cornell University

Ithaca, NY

Undergraduate Teaching Assistant

Aug 2024 – Present

ECE 4750	Computer Architecture	Fall 2026
CS 4850	Probability, Vectors & Matrices in Computing	Spring 2026
ECE 2300	Digital Logic & Computer Organization	Fall 2025
CS 2800	Discrete Structures (Head TA)	Spring 2025
CS 2800	Discrete Structures	Fall 2024

- Designed grading rubrics and helped write problems for classes of 30 to 350 students.
- Led sections, labs, and review sessions, held office hours, and graded assignments.

INDUSTRY EXPERIENCE	NVIDIA Corporation	Santa Clara, CA
	<i>Deep Learning Performance Architect Intern</i>	May – Aug 2026
	- Uncovered new power and clock frequency behavior on the latest GPUs by characterizing how computation and communication interact when overlapped in LLM workloads. - Applied these findings to GEMM and overlap optimizations improving throughput by up to 10% and 15%, and presented an updated power-performance model to the organization.	
	NVIDIA Corporation	Santa Clara, CA
	<i>Deep Learning Performance Architect Intern</i>	May – Aug 2025
	- Built energy monitoring into TensorRT-LLM and Megatron-LM, used to train Nemotron, and implemented application-level DVFS to improve training energy efficiency. - Characterized GPU power and DVFS response for LLM training and inference; my team presented my microsecond-scale power visualizations in a NeurIPS 2025 tutorial.	
	NVIDIA Corporation	Santa Clara, CA
	<i>Systems Software Engineering Intern</i>	May – Aug 2024
	- Set up and ran LLM training and inference on up to 576 GPUs, characterizing their impact on the datacenter power grid to guide rack design and datacenter provisioning. - Ran proxies of large-scale training and inference on a 4-node system, sweeping workload and parallelism parameters to see how they shape utilization, power, and clocks over time.	
PRESENTATIONS	“Characterizing Genomic Workloads on a 1024-Core RISC-V Manycore” <i>NSF PPOSS Panorama Project Meeting, with Po Jui Shih and Niklas Schmelzle</i>	
	<i>Apr 2026</i>	
	“Sorting on the HammerBlade Manycore” <i>NSF PPOSS Panorama Project Meeting</i>	
	<i>Apr 2025</i>	
SELECTED PROJECTS	Full-Custom 8-bit Processor <i>ECE 4740, Cadence Virtuoso</i>	Jan – May 2026
	- Designed a 500 MHz, 229 μW, 1980 μm² Tiny²RV (RISC-V subset) processor in GPD45 with a five-person team, laying out its datapath, control FSM, and register file by hand. - Wrote a simulated-annealing floorplanner to reduce routing congestion, and a solver over all 40,320 instruction encodings to minimize control-unit complexity.	
	ASIC Backend for Allo <i>ECE 6775, MLIR, Google XLS</i>	Oct – Dec 2025
SKILLS	- Extended the Allo accelerator compiler with a Google XLS backend, translating Allo’s MLIR programs into DSLX for synthesis to ASIC RTL. - Implemented lowering for loops, control flow, and memory, carrying matrix-multiply, matrix-vector, vector-add, dot-product, and factorial kernels end-to-end to RTL.	
	<i>Languages:</i>	Python, C/C++, CUDA, Verilog, OCaml, Java, Bash, RISC-V assembly
	<i>Hardware Design:</i>	Cadence Virtuoso, Vivado HLS, Verilator, PyMTL3, Google XLS
	<i>ML & GPU:</i>	PyTorch, Megatron-LM, TensorRT-LLM, Nsight Systems
	<i>Compilers & Tools:</i>	MLIR, Linux, Git, Docker, Slurm