James Contini

(415) - 871 - 4971 | jamescontini@gmail.com | linkedin.com/in/james-contini-3957a79a/ | github.com/XXjcontiniXX

Education

University of California Santa Cruz Santa Cruz, California
Bachelor of Science: Computer Science *Expected June 2025*

- GPA: 3.71
- Relevant Coursework: Compiler Design, Deep Learning, Computer Architecture, Computer System Design
- Activities: Track and Field, Heterogeneous Programming Lab

Research

Honors Thesis March 2025 – June 2025
ScanBox: Tuning Portable GPU Prefix-Scans in Vulkan and WebGPU *Santa Cruz, California*

- Researched and developed kernel engineering strategies advancing performance portability of GPU prefix-scan.
- Complete thesis can be found on website jamescontini.com.

GPU Software Engineer Research Assistant Dec 2023 – Present
Concurrency and Heterogeneous Programming Lab *Santa Cruz, California*

- Identified and described a subgroupBarrier bug in NVIDIA's Vulkan implementation. The issue was rapidly acknowledged and patched in Windows 553.22.
- Reimplemented our lab's open-source Vulkan simplification tool's memory management to use device local buffers enabling accurate GPU benchmarking.

Projects

LLAMA.CPP | *WGSL/WebGPU, C++* Sept 2025 – Present

- Contributing to Llama.cpp's open-source LLM inference library by writing WGSL shaders for their WebGPU backend.

WebGPU Kernel Characterization | *WGSL/WebGPU, C++, Javascript* Dec 2024 – Present

- Designed high performance WGSL prefix-scan shaders inspired by my previous OpenCL kernel designs.
- Iteratively designed implementation using in browser performance benchmarking to fine tune WGSL prefix-sum implementation for peak throughput.

Vulkan/OpenCL Kernel Development | *OpenCL, Vulkan C++, Metal, SPIR-V* Mar 2024 – Dec 2024

- Vulkan prefix-sum kernel achieves up to 43% higher throughput than Nvidia CUB's prefix-sum on small inputs.
- Achieves performance within 1% of Nvidia CUB on RTX 4070 and 1.5% on AMD 7900 XT, relative to device throughput limits.

Computer Vision AI Model | *Python, PyTorch* Nov 2024

- Trained a PyTorch-based Mask R-CNN (Neural Network) for object detection in GPU accelerated HPC system (Jetstream2 @ INDY SCC '24).
- Achieved accurate detection of target images and resolved compatibility challenges within an HPC environment.

Community & Leadership

Sprints Captain Sep 2023 – Present
Track and Field *Santa Cruz, California*

- Achieved fastest 100m in UC Santa Cruz T&F history (2025 - 10.66s)

Technical Skills

Languages: C, C++, Python, OpenCL, JS, CSS, HTML, Haskell, Bash
Frameworks: Vulkan, WebGPU, CUDA, Metal
Developer Tools: VSCode, Ubuntu/Linux, Git, Figma
Libraries: Matplotlib, YACC, Pandas
Applications: Sony Vegas Pro, Fusion360, Slack