Jay Shah

Sharon, Massachusetts | (937) 414-7823 | Email | LinkedIn | Website

Education

University of Dayton Dayton, OH
Master of Science in Computer Science *Aug 2023 – May 2025*
Relevant Coursework: Artificial Intelligence, Data Structure and Algorithms, Data Analysis, Machine Learning

Government Engineering College India
Bachelor of Electronics and Telecommunication Engineering *Aug 2017 – May 2021*
Relevant Coursework: Power Systems, Control Systems, Analog Circuit Design, Signals and Systems

Experience

GPU Developer (WebGPU/WGSL) Sept 2025 – Present
Gridwise *Davis, CA*

- Built and optimized GPU kernels for scan, reduction, radix sort, and histogram using WebGPU, following CUDA-style execution patterns with thread groups, shared memory, and barriers
- Profiled kernels with WebGPU tooling, identifying bandwidth bottlenecks and atomic contention, and tuned memory layouts to achieve sustained throughput of 80 GB/s, improving performance by 40 GB/s over baseline
- Unified multiple histogram implementations into a single, easy-to-use GPU API, simplifying integration for other developers without sacrificing performance
- Performed end-to-end performance testing across workloads, analyzing occupancy, synchronization costs, and memory efficiency, iterating on kernels to maximize throughput
- Focused on practical GPU performance engineering: reduced unnecessary memory transfers, optimized thread-level parallelism, and balanced workload distribution for consistent high performance

Software Engineer Intern Jan 2021 – June 2021
Amar InfoTech *Ahmedabad, India*

- Developed a multi-sensor IoT monitoring system on STM32 using C++, integrating temperature/humidity, air quality and IMU sensors for real-time environmental data acquisition and analysis
- Designed and optimized a real-time data logging and alerting pipeline on STM32, triggering threshold-based notifications for anomalies such as gas leaks, temperature spikes, and abnormal vibration events

Projects

WebSight – Real-Time GPU Profiler | *WebGPU, WGSL, JavaScript* [Live Demo] 2025

- Developed a GPU profiler that tracks compute shader execution, memory usage, and workgroup efficiency without modifying application code, enabling identification of performance bottlenecks
- Implemented memory leak detection, shader complexity checks, and workgroup optimization analysis, providing actionable insights to improve GPU throughput and stability

GPU-Accelerated Medical Imaging Optimization | *CUDA, TensorRT,Nsight* August 2025

- Accelerated a pretrained chest X-ray pneumonia detection model using NVIDIA A100 GPUs, achieving a 50% reduction in end-to-end inference time while maintaining model accuracy
- Applied FP16 precision, layer-wise quantization, and operator fusion across GPU kernels, and used NVIDIA Nsight and TensorRT to profile and eliminate bottlenecks, improving memory efficiency, throughput, and stability

GridDB – GPU-Accelerated Analytics Engine | *WebGPU, CUDA, SQL* [Live Demo] 2025

- Built a GPU-accelerated analytics platform for multi-million record datasets using CUDA and WebGPU, implementing parallel SQL operations (SELECT, WHERE, GROUP BY, ORDER BY, histogram) for high-performance analytics
- Developed and optimized CUDA kernels for aggregation and histogram computations, profiling both WebGPU and CUDA execution, synchronization to maximize throughput and minimize latency
- Produced scatter plots, histograms, time-series charts, and KPI dashboards directly from GPU-processed results, enabling fast insights for quality and process analytics

Technical Skills

Programming & Languages: C/C++, Python, CUDA, WGSL, JavaScript, Embedded C
Frameworks & Libraries: TensorFlow, PyTorch, ONNX, TensorRT, WebGPU, STM32CubeIDE
GPU & Performance Tools: NVIDIA Nsight, TensorRT, Triton, Slurm
Parallel Computing: CUDA, OpenMP, MPI
Operating Systems: Linux