Avery Elliott

512-555-7306 avery.elliott@example.com
linkedin.com/example/averyelliott Austin, TX 78701

SUMMARY

Senior GPU performance software engineer with eight years of professional C++ development and more than five years focused on accelerator kernels, performance libraries, and low level optimization. Builds production quality GEMM, convolution, attention, reduction, JIT, and **code generation** infrastructure across modern GPU architectures. Delivers **mixed precision** and **quantized paths** using BF16, FP16, INT8, FP8, and FP4, supported by profiling, benchmarking, analytical models, and **memory traffic** analysis. Brings practical depth in SYCL, DPC++, CUDA, OpenCL, SIMD, OpenMP, oneTBB, caches, memory systems, multithreading, and parallel algorithms. Partners effectively with compiler, architecture, validation, and framework teams across PyTorch, TensorFlow, OpenVINO, and ONNX Runtime. Ready to apply this experience toward dependable, high impact performance software.

EXPERIENCE

Senior GPU Performance Software Engineer

March 2023 - Present

Aurora Parallel Systems

Austin, TX

Owns GPU math libraries, JIT code generation, mixed precision paths, profiling, validation, and accelerator co design. Leads delivery across performance, compiler, architecture, and framework partners, converting low level improvements into dependable production infrastructure.

- Led GEMM and **convolution** kernels with CUDA and SYCL, improving throughput across modern GPU architectures.
- Built JIT code generation paths for fused **attention** and **reduction** primitives, reducing runtime dispatch overhead.
- Profiled memory traffic and cache behavior, removing bottlenecks from production GPU execution paths.
- Co designed accelerator primitives with **compiler teams**, aligning hardware capabilities with practical library interfaces.
- Expanded BF16, FP16, INT8, FP8, and FP4 validation through benchmarking and continuous integration.
- Connected optimized primitives with PyTorch, **TensorFlow**, OpenVINO, and **ONNX Runtime** integration tests.

GPU Software Engineer

July 2020 - February 2023

Summit Compute Labs

Boulder, CO

Developed GPU kernels and performance tooling across CUDA and SYCL or DPC++ environments. Supported quantization, framework validation, benchmarking, and runtime analysis, giving product teams clearer evidence for architecture and delivery decisions.

- Optimized **CUDA GEMM** kernels through tiling and shared memory analysis, raising sustained workload performance.
- Ported selected primitives to SYCL and DPC++, preserving correctness across accelerator backends.
- Built quantized **INT8** and mixed precision paths, strengthening inference coverage for framework partners.
- Created benchmark suites for convolution, attention, and reduction kernels, exposing regressions before release.
- Validated runtime behavior across **PyTorch**, TensorFlow, and ONNX Runtime integration environments.
- Partnered with **validation** engineers to convert **profiling** findings into repeatable CI checks.

Performance Software Engineer

June 2018 - June 2020

Redwood Systems Software

Portland, OR

Built modern C++ performance libraries for CPU and parallel workloads, with emphasis on **SIMD**, cache behavior, **multithreading**, and repeatable profiling. Delivered foundational optimization experience that transferred directly into later GPU kernel and runtime work.

- Implemented **modern C++** math library routines with SIMD **vectorization**, improving throughput on compute intensive workloads.
- Parallelized hot paths with **OpenMP** and oneTBB, shortening execution time across multicore systems.
- Analyzed cache locality and **memory systems**, guiding data layout changes for steadier performance.
- Built profiling workflows that connected hardware counters with actionable optimization decisions.
- Added benchmark coverage for regression detection, giving reviewers clear evidence before library releases.
- Documented optimization tradeoffs for software and hardware stakeholders, improving adoption of performance changes.

PROJECTS

Mixed Precision Kernel Benchmark Suite 2024

Small measurement choices reveal large runtime effects. Built a cross platform benchmark suite comparing GEMM, convolution, attention, and reduction kernels across BF16, FP16, INT8, FP8, and FP4 paths.

Portable JIT Primitive Runtime 2022

Fast code starts with useful choices. Developed JIT and code generation infrastructure supporting hardware aware dispatch, fusion experiments, profiling hooks, and validation across CUDA and SYCL environments.

LEADERSHIP & AWARDS

- Received 2024 Performance Engineering Recognition for GPU runtime improvements and cross team technical leadership.
- Earned 2022 Kernel Optimization Award for advancing portable CUDA and SYCL primitive implementations.
- Led architecture and compiler working sessions that strengthened shared performance decisions across engineering teams.

EDUCATION

Master of Science in Computer Engineering 2018

Oregon State University GPA: 4.0 *Corvallis, OR*

Coursework: *Computer architecture, parallel computing, GPU programming, performance optimization*

Bachelor of Science in Computer Science 2016

University of Nevada, Reno GPA: 4.0 *Reno, NV*

Coursework: *Data structures, algorithms, operating systems, computer organization*

CERTIFICATIONS

- CUDA C/C++ Accelerated Computing 2021
- oneAPI Developer Certification 2024

TECHNICAL SKILLS

- **GPU Programming:** CUDA, SYCL, DPC++, OpenCL
- **GPU Kernels:** GEMM, convolution, attention, reductions
- **Precision Formats:** BF16, FP16, FP8, FP4
- **Quantization:** INT8, quantized paths, mixed precision
- **C++ Development:** Modern C++, templates, standard library
- **Parallel Computing:** OpenMP, oneTBB, multithreading
- **Vectorization:** SIMD, vector instructions, data parallelism
- **Performance Analysis:** Profiling, benchmarking, performance models
- **Memory Systems:** Caches, memory traffic, data locality
- **Framework Runtime:** PyTorch, TensorFlow, OpenVINO
- **Inference Runtime:** ONNX Runtime, dispatch, validation
- **Compiler Infrastructure:** JIT, code generation, compiler tuning

SKILLS

- Modern C++
- GPU kernels
- SYCL and DPC++
- CUDA
- OpenCL
- GEMM
- Convolution
- Attention
- Reductions
- JIT code generation
- Mixed precision
- Quantization
- Profiling
- SIMD vectorization
- Parallel programming

PROFESSIONAL AFFILIATIONS

- Member, Association for Computing Machinery, active in systems and performance engineering communities.
- Member, Khronos Group community, following open standards for heterogeneous and accelerator computing.

LANGUAGES

- English (Native)
- Spanish (Intermediate)

ADDITIONAL INFORMATION

Work Status : Authorized to work in United States. No sponsorship required.

REFERENCES

AVAILABLE ON REQUEST