Logan Chu

Kirkland, WA | (425) 241-8158 | lc478@duke.edu | linkedin.com/in/loganchu | github.com/LoganChu

Education

Duke University Durham, NC

Bachelor of Science in Computer Science; GPA: ***3.92*** *May 2027*

- **Relevant Coursework:** Machine Learning (ML) Systems, ML & Deep Learning, Mathematics of ML, Distributed Systems, Computer Networks, Operating Systems, Computer Architecture, Digital Systems, Data Structures & Algorithms

Technical Experience

Software Engineering Intern, Yotta Labs December 2025 – Present

- Profiled **vLLM** and **SGLang** inference stacks on open-source LLMs (LLaMA, Qwen) across batch sizes, isolating latency/throughput bottlenecks from **KV-cache** growth, kernel fusion, and scheduler behavior.
- Evaluated NVIDIA Nsight vs AMD ROCm GPU profilers at kernel and memory-subsystem granularity, identifying tooling gaps that constrained occupancy analysis, memory tracing, and cross-vendor optimization.

Tech Director, Campus Enterprises (Student-Run Company) Feb 2025 – Present

- Increased marketing efficiency **40%** with an **AWS Bedrock Agent** for natural language querying of **PostgreSQL** data.
- Designed a distributed **Apache Spark** workflow to load relational data into Redshift for retrieval-augmented generation.
- Enabled **17K+ users** by engineering scalable, serverless AWS Lambda **APIs** (Node.js) with VPC isolation, reducing infrastructure costs by **70%** and optimizing performance for ML-driven microservices.

Research Assistant, Grill Lab Apr 2025 – Sep 2025

- Improved fascicle segmentation accuracy **15%** by training a U-Net model in **PyTorch** for a **$16M** NIH vagus nerve project.
- Boosted data diversity **23%** by deploying custom CVAT AI **Docker** container for high-res image annotation.
- Reduced training time 5 fold by **parallelizing** fold training with **SLURM** on the Duke Computer Cluster (Linux-based).

Projects

Fused GPU Inference Kernels Oct 2025 - Present

- Implemented a custom Triton GPU kernel fusing GEMM, bias add, and ReLU with **shared-memory tiling** and FP16 register accumulation, delivering **30% runtime improvement** over PyTorch eager execution on T4 GPUs.
- Optimized GPU memory hierarchy and execution by eliminating intermediate writes via **operator fusion** and tuned tile sizes, **reducing global memory traffic** and kernel launches across 5,000-iteration latency-critical benchmarks.

ML Graph Runtime Oct 2025 - Present

- Built a TensorFlow-v1–style **automatic differentiation** runtime with topologically sorted graph execution and vectorized ops, enabling end-to-end Transformer training with 50 % MNIST accuracy.
- Implemented fused MatMul+Softmax and MatMul+LayerNorm operators with correct backward graphs, cutting intermediate memory traffic and achieving **1.5× throughput** over unfused baselines.

HPC Agent for Duke Compute Cluster Sep 2025 – Present

- Built a terminal-based LLM agent for the Duke Compute Cluster using **LiteLLM** and an **MCP** server, deploying stateless services on **Kubernetes** to support **20+ concurrent users** with reproducible infrastructure via **Bazel** and **Ansible**.
- Enabled 100% auditable human-in-the-loop HPC recommendations by architecting safety-first LLM infrastructure with OPA policy enforcement and OpenTelemetry observability across a production inference pipeline.

Fast Diffusion Inference via Probabilistic Step-Skipping Aug 2025 - Dec 2025

- Engineering a stochastic step-skipping scheduler for lightweight (2-10M parameters) **diffusion models**, reducing inference latency **30–50%** while preserving image fidelity (FID/IS) through probabilistic timestep omission.
- Benchmarking **stochastic sampling** against DDPM/DDIM baselines, quantifying efficiency–fidelity trade-offs.

Skills

Inference Systems: PyTorch, vLLM, SGLang, TensorRT-LLM, Triton, KV Cache, Continuous Batching, Quantization

Inference Optimization: Kernel Fusion, Graph Optimization (TorchCompile, XLA), Profiling (Nsight), CUDA Streams, Mixed Precision (FP16, BF16)

Infrastructure: AWS, GCP, Docker, Kubernetes, MLflow, Bazel, Git, CI/CD, Linux

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