Calvin Huang

calvinh99.github.io
jibunhidoi999@gmail.com | (530) 631-9585 | 2130 Greer Rd, Palo Alto, CA

EDUCATION

STEVENS INSTITUTE OF TECHNOLOGY

BS in Computer Science
Jan 2024 | Hoboken, NJ

LINKS

Github:// **calvinh99**
LinkedIn:// **calvinh99**
Twitter:// **@scholarc1314**

COURSEWORK

GRADUATE

Deep Learning
Statistical Machine Learning
Natural Language Processing
Computer Vision
Parallel & Distributed Computing

UNDERGRADUATE

Systems Programming
Compilers & Language Design
Computer Architecture
Data Structures & Algorithms
Theory of Computation
Multivariate Calculus & Optimization
Linear Algebra

SKILLS

LANGUAGES

Python • C++ • CUDA
Bash • C

ML FRAMEWORKS

JAX • PyTorch • XLA
vLLM • TensorRT-LLM
HuggingFace Transformers

COMPILER & RUNTIME

Triton • CUDA Runtime
NCCL • cuDNN
NVIDIA Nsight • xprof

DISTRIBUTED TRAINING

FSDP • DeepSpeed
Tensor/Pipeline Parallelism
Gradient Checkpointing

INFRASTRUCTURE

Linux • Docker • AWS
A100/H100 GPUs • Slurm

EXPERIENCE

AMAZON | Rufus LLM - ML Systems Engineer

March 2024 – Present | Palo Alto, CA

- Profiled and optimized distributed LLM training across multi-node GPU clusters using NCCL collectives, reducing all-reduce communication overhead by 35% through gradient bucketing and overlap with compute.
- Debugged XLA compilation failures and memory OOM issues using HLO graph analysis; implemented custom sharding annotations to reduce peak memory by 40% on 70B parameter models.
- Built distributed profiling tooling to trace per-device memory consumption and kernel execution timelines, identifying fusion opportunities that improved training step time by 18%.
- Optimized inference serving with continuous batching and KV-cache memory management in vLLM; reduced P99 latency by 45% while increasing throughput 3x on A100 clusters.

LINKDEA | ML Infrastructure Engineer

May 2023 – Feb 2024 | San Jose, CA

- Migrated training pipeline from PyTorch to JAX/Flax, implementing custom pjit sharding strategies (FSDP + tensor parallelism) that scaled Llama2-13B training to 64 A100 GPUs with 92% compute efficiency.
- Wrote custom Triton kernels for fused attention and RMSNorm, achieving 2.3x speedup over cuDNN baselines by optimizing shared memory tiling and register allocation.
- Integrated TensorRT-LLM for production inference; implemented INT8 weight-only quantization and speculative decoding, reducing latency by 4x while maintaining <1% quality degradation.
- Built memory profiling tools to track JAX buffer allocations across training steps, identifying rematerialization opportunities that reduced HBM usage by 28%.

STEVENS INSTITUTE FOR ARTIFICIAL INTELLIGENCE | Research Assistant

June 2023 – Sep 2023 | Hoboken, NJ

- Optimized Stable Diffusion inference by analyzing XLA fusion patterns and implementing custom scheduling for UNet attention blocks, reducing sampling latency by 2.1x on single GPU.

PROJECTS

C4CROW | AlphaZero in JAX

Implemented AlphaZero from scratch in JAX with jit-compiled MCTS and vectorized self-play. Achieved 50x speedup over naive Python through XLA fusion of tree search operations.

SIMPLE-BACKPROP | Custom Autograd Engine

Built neural network framework from scratch implementing reverse-mode autodiff. Manual gradient computation for Dense, ReLU, Softmax, Cross-Entropy with memory-efficient backward pass.

MANGASTYLE | Full-Stack ML Platform

Built art platform serving manga-style images—AWS EC2, Django, NGINX, MySQL, Docker with async task queues.