Chongfeng Ling

An AI Infrastructure engineer deeply focused on low-level model optimization.

career.ling@outlook.com | chongfengling.github.io | Chongfeng Ling | chongfengling

Summary

AI Infrastructure Engineer specializing in low-level Large Language Model (LLM) optimization. Proficient in CUDA and Ascend C kernel development and hardware architecture. Dedicated to pushing hardware limits for maximum computational performance.

Experience

FESCO Adecco Co., Ltd. **Nov. 2024 – Oct. 2025**
AI Kernel Engineer (Ascend C) *Shanghai, China*

Quantized MatMul Kernel: Ascend-based Quantized MatMul Kernel Development for **DeepSeek MoE**.

- Spearheaded the **W8A8/W8A4 full-quantization** pipeline. Implemented fine-grained quantization (**Per-Token/Group/Channel**) to compress memory footprint while preserving long-context generation accuracy.
- Implemented advanced **Split-K** strategies to maximize **Cube throughput**. Re-architected the **Fixpipe pipeline** to fuse **Dequantization** in local memory, leveraging **Double Buffering** to completely overlap **DMA latency**.
- Reduced memory footprint by ∼**50%**, boosted operator throughput by **1.5x-1.8x** compared to FP16 baselines, and reduced end-to-end latency by over **30%**, enabling ultra-efficient and low-cost serving for DeepSeek models.

Projects

Memory-Aware GEMM & SpMV Engine | *cuda, numba, python* [repo]

- Developed highly parallel Numba CUDA kernels for PDEs. Benchmarked **C/Fortran-contiguous layouts** in GEMM, exploiting optimal combinations for inner-product alignment. It guaranteed **successive memory reads**, maximizing **SIMD acceleration** and **cache coalescing** for an **30x** speedup over naive implementations.
- Engineered a custom **CSR format** and JIT-compiled **SpMV operator** integrated into GMRES solvers. Bypassed Numba limits via JIT-wrappers for parallel **COO assembly**. Profiled FEM systems to identify strict **sequential dependencies** in MINRES, establishing theoretical parallel scaling limits.

High-Performance N-Body Simulation | *C++, OpenMP, MPI* [repo]

- Engineered a **high-performance planetary simulation system** using C++, achieving a 50x speedup through strategic compiler optimizations (-O2) and memory management.
- Implemented multi-core parallelism using **OpenMP**, conducting rigorous hard and weak **scaling experiments** to analyze thread overhead and synchronization bottlenecks.

CUDA-Handson | *cuda, ncu, nsys* [repo]

- Iterated 7 versions of **SGEMM** kernels, exploiting Shared/Register Tiling, vectorized memory access (float4), and Padding to eliminate Bank Conflicts, leveraging WMMA (Tensor Cores) to achieve peak performance at **109% of cuBLAS**, while building an end-to-end FP32-to-INT8 **quantization** pipeline (MSE $<$ 1.8e-3).
- Overcame memory-bound bottlenecks in Softmax by implementing lock-free **Warp-level** reductions and the **Online Softmax** (Single-pass) algorithm to drastically reduce global memory traffic, utilizing Nsight Compute to deeply profile and resolve microarchitectural trade-offs between register pressure and hardware occupancy.
- Utilized **nsys** to trace Host-Device interactions and CUDA stream scheduling, maximizing computation-DMA **transfer overlap**. Leveraged **ncu** to deep-dive into microarchitectural bottlenecks and quantitatively optimize SM **Occupancy and Memory Throughput**.

Other Open-Source Projects

- **Conway's Game of Life**: Object-oriented simulator design for the investigation of stationary patterns. [repo]
- **Supervised Learning**: Kernel Ridge Regression and multiclass kernel perceptrons for predictive modeling. [repo]
- **Awesome-XJTLU**: An open-source collaborative knowledge platform tailored for undergraduate students. [repo]
- **Hydrological Data Toolkit**: Collaborative Python toolkit designed for temporal data analysis and visualization.

Education

University College London **Sep. 2022 – Dec. 2023**
Master of Science in Scientific and Data Intensive Computing *London, UK*

University of Liverpool **Jun. 2022**
Bachelor of Science in Applied Mathematics *Liverpool, UK*

Xi'an Jiaotong-Liverpool University **Sep. 2017 – Jun. 2022**
Bachelor of Science in Applied Mathematics *Jiangsu, China*

Skills

Languages : Cpp (CUDA C, Ascend C), Python (PyTorch)
Frameworks/Tools : Git, Linux, NVIDIA Nsight Systems (nsys), Nsight Compute (ncu)

Last updated on 2026-04