Yibai Meng

Mountain View, CA ✉ yibai@meng.engineer linkedin.com/in/yibai-meng ☎ +1 6692009099

Skills

- **Programming Languages**: Python, C/C++, Go
- **Machine Learning**: PyTorch, Jax, XLA, MLIR, Transformer/LLM, quantization, ML model optimization
- **GPU**: CUDA, OpenAI Triton, PTX, Nsight Compute, CUTLASS

Industry Experiences

Waymo Jan 2023 - Present
Software Engineer *Mountain View, California*

- ML performance engineer, focusing on ML model optimization and inference infrastructure.
- **GPU Kernel Development**: Implemented high-performance custom GPU kernels, including flash attention, quantized feed-forward operations, fused convolution, and Waymo-specialized operations in OpenAI Triton, CUDA and inline PTX. Profiled and fine-tuned these kernels using Nsight Compute, identifying subtle code generation issues within the underlying stack. Developed kernels across multiple precisions, including INT8 and FP16. Conducted in-depth analysis of generated PTX code and achieved bit-for-bit accuracy with the JAX XLA reference implementation. Leveraged PTX assembly and CUDA intrinsics for low-level optimizations. These efforts led to a 3× end-to-end speedup over XLA for a critical VLM model, enabling on-vehicle deployment. Also implemented a novel sparse activation approach in CUDA, utilizing raw CUDA primitives, leading to 2x speedup.
- **LLM Quantization**: Implemented quantization for transformer kernels, including 8-bit weight-and-activation, 4-bit weight-only, and 4-bit weight-and-activation schemes. For 4-bit weight-only quantization, applied bitwise operations and inline assembly to overcome the lack of native Triton support. For 4-bit weight-and-activation quantization, leveraged CUTLASS to develop a high-performance fused quantized projection kernel.
- **Model Optimization**: Improved the latency and stability of on-vehicle machine learning models through quantization and operation fusion. One fusion optimization reduced latency by 30% compared to XLA. Designed a bespoke fused convolution module with a custom quantization scheme and kernel, enabling model scalability. Modernized graph manipulation workflows using MLIR.
- **Infra and Tooling**: Designed and implemented the custom kernels framework used across Waymo; developed continuous integration testing and benchmarking infrastructure; created a tool for inspecting models and providing optimization suggestions; and built a tool to verify numerics after graph manipulation, leveraging the existing integration testing infrastructure.

TikTok May 2022 - Aug 2022
Software Engineer Intern *Mountain View, California*

Worked on software defined network, implemented a novel data plane network verification algorithm in C++ from scratch.

Education

University of California, Berkeley Aug 2021 - Dec 2022
Master of Engineering in Electrical Engineering and Computer Science *Berkeley, California*

Peking University Sep 2016 - May 2020
Bachelor of Science in Electronics and Information Science and Technology *Beijing, China*

Academic Experiences

Center for Energy-Efficient Computing and Applications, Peking University July 2020 -- June 2021
Research Assistant *Beijing, China*

Implemented GPU acceleration of elfPlace, a nonlinear, nonconvex optimization algorithm for FPGA physical synthesis. Reframed the optimization problem as a neural network training task and used PyTorch C++ with CUDA extensions to optimize critical segments. Achieved an average runtime reduction of 7×. Resulted in two academic publications in top journals.

Publications

- **elfPlace: Electrostatics-based Placement for Large-Scale Heterogeneous FPGAs**: Yibai Meng, Wuxi Li, Yibo Lin and David Z. Pan. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD)*, 2021
- **Multi-Electrostatic FPGA Placement Considering SLICEL-SLICEM Heterogeneity and Clock Feasibility**: Jing Mai, Yibai Meng, Zhixiong Di and Yibo Lin. *Design Automation Conference (DAC)*, 2022

Updated 2025-08-25, see https://www.mengyibai.com/cv.pdf for current version.