Yibai Meng

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

Skills

Programming Languages: Python, C/C++, Go, JavaScript/HTML

Machine Learning and GPU: PyTorch, Tensorflow, Jax, CUDA, XLA, MLIR, Triton, quantization

Industry Experiences

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

- Working on ML model optimization and inference infrastructure.
- **GPU Kernel development**: Implemented high performance transformer kernels (flash attention, quantized feed-forward) in OpenAI Triton. Profiled and fine-tuned these kernels using Nsight Compute, identifying subtle code-gen issues in our underlaying stack. Implemented low-bit quantization while maintaining model quality. Achieved 3x end to end speedup compared to XLA, unblocking critical foundation model for on-car usage. Designed and created continuous intergration testing and benchmarking infrastructure.
- **Model optimization**: Improved the latency and stability of on-vehicle ML models using techniques such as quantization and non-trival operation fusion. One such operation fusion lead to 30 percent latency reduction compared to XLA. Modernized graph manipulation workflows with technologies like MLIR.
- **Foundation model support**: Unblocked foundation model adoption by overcoming longstanding limitations due to tensorflow implementation details. Made numerous contributions to tensorflow codebase.
- **Developer tooling**: Created tool for inspecting models and providing optimization suggestions. Made tool to verify numerics after graph manipulation, leveraging existing integration testing infrastructure.

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

- As a member of the Software Defined Network team, implemented a novel data plane network verification algorithm in C++ from scratch, based on a recent academic paper.
- Drastically increased the performance of network topology modelling and invariant checks (loop) by 500+ times, compared with existing Python tools, enabling real-time verification of TikTok's expansive global network.

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*

- GPU acceleration of elfPlace using CUDA, an algorithm for the placement phase of FPGA physical synthesis. Shortened runtime by 7 times on average.
- Ported the algorithm to an PyTorch based framework, framing this nonlinear nonconvex optimization problem as training a neural network. Used C++ & CUDA extension on Python to speed up critical segments while maintaining low code complexity. Refactored the individual functionalities into "operators", following the paradigm of high-cohesion low-coupling.
- Extended the algorithm to consider clock network routing resource constraints with a quadratic penalty, with consideration of both global placement convergence and design legality.
- Extended the algorithm to include new cell types, allowing it to process a non-academic real-life industry architecture and substantially improved its performance.
- 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