Simon Boehm

github.com/siboehm Technical Blog: siboehm.com

Open Source Projects

- **Decision tree compiler (lleaves)** : An LLVM compiler for tree-based ML models. Speeds up prediction by ∼10x through cache blocking and profile-guided optimization. Used in production at QuantCo. Sole developer & maintainer, >50K downloads on PyPI. Featured in academic database literature .
- **Tech blog (siboehm.com)** : 10K monthly visitors. Notable posts: How to Optimize a CUDA Matmul Kernel for cuBLAS-like Performance , Can Function Inlining Affect Floating Point Consistency? , Pipeline-Parallel Distributed Training .
- **Distributed deep learning (ShallowSpeed)** : Implements distributed training of sequential deep learning models, built on Numpy and MPI. Supports interleaved data parallelism and pipeline parallelism (GPipe). Also see accompanying blog post .
- **Conditional density estimation (cde)** : TensorFlow library implementing ML algorithms for density estimation. Contributed normalizing flows and variational inference. Paper available on ArXiv , >80K downloads on PyPI.

Experience

- **Astera Institute** Oct. 2022 - present
  HPC Software Engineer (Computational neuroscience team)
  - **Research Engineer**: Scaled up Axon, a research codebase that models the mammal brain including spiking and calcium-based learning. Added fine-grained multithreading and MPI distributed training.
  - **CUDA Rewrite**: Completed rewrite of existing 20K LOC Golang research codebase for Nvidia GPUs. Implemented kernels that leverage sparse spiking. Debugged emergent behaviours of large models during training.
- **AMD** Feb. 2022 - July 2022
  Compiler Engineering Intern (Deep Learning compiler team)
  - **MLIR compiler bringup**: Implemented estimation of compute cycles and bandwidth demand for L1-fused neural network kernels. Worked on mapping of >10B parameter models onto AIE accelerator cards.
  - **PyTorch internals**: Contributed to compiler-frontend, lowering operators from PyTorch ATen kernels to torch-mlir and ONNX. Focus lied on representing quantization for transformer and vision models.
- **QuantCo** Oct. 2020 - Dec. 2021
  Data Engineering Intern (Insurance pricing team)
  - **Analytics platform speedups**: Worked on the internal analytics platform that is used by >25 data scientists to develop models. Migrated model checkpoints to SquashFS for 4x faster model reuse and 3x disk space savings.
  - **Datastore migration**: Migrated the main internal datastore from on-disk storage to Google Cloud Storage. Resulted in a new open source package (minimalkv).
- **First Momentum Ventures** June 2018 - Sep. 2020
  Tech Associate (early-stage VC)
  - **Startup database**: Together with CTO developed a detailed, auto-updating database of 15K early-stage European startups. System utilized distributed scraping, NLP-based content extraction, and data cleaning.
  - **Product development**: Worked on Django backend for using startup database to semi-automatically find new investments. API database access was sold to stakeholders, reaching 10K€ MRR.

Education

- **ETH Zürich** Sep. 2019 – Jan. 2022
  Master of Science in Data Science; Swiss grade average: 5.5
  NeurIPS 2022 publication : "Predicting Single-Cell Perturbation Responses for Unseen Drugs" by Simon Böhm*, Leon Hetzel*, et al.
- **Karlsruhe Institute of Technology (KIT)** Oct. 2015 – Sep. 2019
  Bachelor of Science in Computer Science; German grade average: 1.3 (top 5% of class)
  Thesis: "Conditional Density Estimation with Normalizing Flows."
  Awards: Scholarship from the German Academic Scholarship Foundation (Studienstiftung).

Programming Skills

- **Languages:** Python, C++, CUDA, Golang **Technologies:** PyTorch, ONNX, MLIR, Docker