Bangda Zhou

bangda.zhou@gmail.com — **San Francisco Bay Area** — **765 - 413 (6089)**

Experience

Staff Software Engineer, Tesla, Inc — **August** 2023 — **Present**

- Dojo, ML Infrastructure & Performance
  - **Team Leadership:** Lead the Dojo ML Infra & Performance team, enabling Tesla's neural networks to train efficiently on Dojo. Drive large-scale model training and optimize business-critical services.
  - **PyTorch Integration:** Proposed and implemented a distributed backend supporting PyTorch native collectives; designed and added `torch.compile` support.
  - **Training Scalability:** Designed and built a Dojo-optimized FSDP2 wrapper supporting flexible sharding and scaling strategies (DP, TP, EP, PP).
  - **Distributed Graph Compiler:** Developed a graph compiler infrastructure from scratch, including bufferization (static buffer management) and collective scheduling for compute–communication overlap. Built resource management for collectives (buffers, barriers, semaphore, etc.).
  - **Low-Precision Training:** Directed FP8 hardware design explorations; integrated FP8 training into Dojo with custom scaling methods and Dojo-specific FP8 formats.
  - **Correctness Verification:** Created a model-level numerical verification framework to ensure consistency across heterogeneous devices.
  - **Kernel Development:** Supported custom kernel creation and integration into the training stack.
  - **Inference Optimization:** Overhauled the offline inference service to improve compatibility and deliver out-of-box performance gains.
  - **Mentorship:** Onboard and mentor new engineers to accelerate team impact.

Senior Software Engineer, Google, Inc — **March** 2018 — **August** 2023

- ML Model Serving Runtime/Compiler

  Lead the business critical runtime infrastructure for fleetwide machine learning model serving (model inference) from Tensorflow or JAX on TPU:

  - Developed the high performance serving runtime/compiler infrastructure which is adopted by different serving services fleetwide.
  - Improved Large Language Model and other ML model serving performance.
  - Optimizations and graph rewrites based on MLIR compiler stack.
  - Hands-on experience on optimizing ML workload on accelerator (XLA:TPU).
  - Large model partitioning (SPMD)
  - Lead a team to develop proposed features from collaborating with model developers and profiling the production workload.

  For more information, see Published Technical Article, Github Tensorflow Runtime.

Senior R&D Software Engineer II, Synopsys, Inc — **August** 2015 — **March 2018**

Static Timing Analysis. Proposed and implemented algorithms for large-scale transistor-level circuit simulations. Analyzed trade-off among runtime, accuracy, and memory.

Computer Scientist (Intern), Sandia National Laboratory — **May** 2013 — **September** 2013

Electrical Models & Simulation Group. Contributed to project *Trilinos*, an open sourced high performance software framework for solving large-scale complex multi-physics engineering and scientific problems.

Education

Purdue University, West Lafayette, IN., USA

PhD, *Electrical and Computer Engineering*, Advisor: *Prof.* Dan Jiao — **August** 2015

- *Linear Complexity Direct Finite Element Solver*
  Fastest, and first linear (optimal) complexity direct finite-element solver for large-scale engineering analysis, greatly outperforms state-of-the-art direct sparse matrix solvers. [read more here]

Shanghai Jiao Tong University, Shanghai, China

BS, *Electrical Engineering*, — **June** 2010

AWARDS

- Feats of Engineering Award, Google, 2020, 2022
- 16 Spot bonus and 10 peer bonus, Google
- Best Student Paper Award, 2nd Place, ACES, 2015
- Best Paper in Session Award, SRC TECHCON, 2014
- Best Student Paper Finalist Award, IEEE Int'l. Symp. on Antennas and Propagation, 2013

PUBLICATIONS 14 peer-reviewed journal and conference papers.