Xing Liu

Email: xingliu14@gmail.com
Website: https://xingliu14.github.io/
Mobile: +1-346-221-9933

Professional Experience

- **Google** — Mountain View, CA
  Software Engineer — *Sept. 2021 - present*
  - Google Shopping LLM Fleet Serving and Optimization: engaged with, created, and managed Gemini serving configurations for 20+ clients within Google Shopping. Applied various optimizations, including EEVEE MTP, sharding tuning, KV Cache quantization, and XLA flag tuning, to achieve fleet-wide 2000+ TPU savings
  - Gemini 2.0 Flash Quantization: applied int4 weight quantization to FFW layers and int8 weight quantization to ATTN layers to Gemini 2.0 Flash decode serving workload, with sub-channel config, achieved neutral quality and 1.1x speedup
  - StableHLO Quantizer: Lead the design and implementation of a neural network Quantizer (open-source) at StableHLO level. This is a compiler-level quantizer that provides robust cross-framework quantization capabilities with hardware optionality
  - Maintained personal search engine which let multiple Google internal clients to generate data which compliance with the privacy policy
- **Yahoo** — Sunnyvale, CA
  Software Engineer — *Mar. 2019 - Sept. 2021*
  - Comm Channel Reputation Datasets: Implemented a data mining pipeline to generate reputation scores for 1.8B comm channels (email addresses/phone numbers), aggregated comm channel reputation data to produce comm channel groups reputation (email domains/phone carriers), built heuristics set and machine learning model for the pipeline
  - Baltar: Improved and maintained a classifier on Verizon Media accounts platform to identify malicious users using machine learning models with Hive, PySpark, Oozie and HDFS
  - Developed dashboards to display realtime performance metrics for each model and heuristic, enable engineers to monitor and compare differences in performance easily
- **Yahoo** — Sunnyvale, CA
  Software Engineering Intern — *May 2018 - Oct. 2018*
  - MTA-STS: Implemented RFC 'SMTP MTA Strict Transport Security' with Java, integrated the library into Yahoo! mail outbound client and made Yahoo! the second mail service provider to support this mail security protocol
  - Developed alternative dependency service for JIRA with Go, AWS, Ansible and DynamoDB, which significantly improved team pipeline's stability and speed

Education

- **Rice University** — Houston, TX
  Master of Computer Science — *Aug. 2017 – Dec. 2018*
- **Tsinghua University** — Beijing, China
  Bachelor of Engineering in Microelectronics — *Aug. 2012 – July 2016*

Skills

- **Programming**: C++, Java, Go, Python, Unix Shell, SQL, JavaScript, HTML/CSS, Typescript
- **Frameworks and Tools**: MLIR, Hive, PySpark, Oozie, Angular, Git