# 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*
  - Led large-scale LLM optimization projects for Google Shopping, a role that involved creating and managing Gemini serving configurations for a portfolio of over 20 clients. Delivered a 2,000+ TPU reduction in resource consumption by applying sophisticated techniques like EEVEE MTP and KV Cache quantization, directly contributing to faster online inference and higher offline processing throughput
  - Gemini 2.0 Flash Quantization: implemented and applied int4 weight quantization to FFW layers and int8 weight quantization to ATTN layers to Gemini 2.0 Flash decode serving workload at the TPU HLO layer, with sub-channel config, achieved neutral quality and 1.1x speedup
  - StableHLO Quantizer: Lead the design and implementation of a neural network Quantizer (open-sourced) at StableHLO level. This is a compiler-level quantizer that provides robust cross-framework quantization capabilities with hardware optionality
  - Developed and optimized a core server infrastructure for Google Shopping LLM services, enabling advanced features like real-time streaming and a sophisticated caching mechanism. Engineered the system to perform configurable TnS checks, ensuring secure and compliant service delivery
  - uLLM/vLLM: Developed quantization features for the inference engine for TPU which utilize vLLM as a foundation framework
- **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

## Project

- **AI-Powered B2B Hotel Booking Platform** — Dec. 2024 - present
  - Architected and deployed a business travel booking platform that currently serves over 10 companies, supporting 1,000+ MAU and an extensive inventory of thousands of hotels
  - Engineered an AI agent utilizing agentic workflows and RAG pipelines to automate and streamline hotel booking, trip planning, and inquiry responses directly within a chat application

## 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 and Frameworks**: C++, Java, Python, MLIR, Go, SQL