Chijun Sima

simachijun@gmail.com | LinkedIn: chijun-sima

Research Interests

Efficient AI (MLSys): scalable and cost-effective training/serving; system–algorithm co-design.

Education

South China University of Technology Sep 2016 – Jun 2020

B.Eng. in Computer Science and Technology (Innovation Class); GPA: ***3.85/4.00*** *(Rank* ***1/28****)*

Selected Publications

- **Ekko: A Large-Scale Deep Learning Recommender System with Low-Latency Model Update**. **Chijun Sima***, Yao Fu*, Man-Kit Sit, Liyi Guo, Xuri Gong, Feng Lin, Junyu Wu, Yongsheng Li, Haidong Rong, Pierre-Louis Aublin, Luo Mai (*co-first).
  OSDI ’22. Supervised by Luo Mai.
- **Dynamic Barycenter Averaging Kernel in RBF Networks for Time Series Classification**. Kejian Shi, Hongyang Qin, **Chijun Sima**, Sen Li, Lifeng Shen, Qianli Ma. **IEEE Access**, 2019.

Research & Industry Experience

Tencent (WeChat) Jul 2020 – Present

Senior Software Development Engineer — Efficient ML Systems *Guangzhou, China*

Ekko: low-latency model update for multi-terabyte DLRMs (published in part as OSDI’22)

- **Problem.** Scaling DLRMs improved offline accuracy but degraded online engagement; diagnosed the root cause as **stale models** from increased **model-update latency** under pre-scaling infrastructure.
- **Key idea.** Co-designed deployment mechanisms with **model-aware** policies to maintain second-level freshness at extreme scale.
- **Technical contributions.**
  - **Update dissemination + scheduling:** designed compressed update dissemination and an accuracy-aware scheduler prioritizing updates using gradient/model signals; reduced WAN bandwidth by **92%**.
  - **SLO-aware placement:** built a shard manager using mathematical optimization to co-locate models without burdening inference engines; reduced machine cost by **49%**.
  - **Safe rollout:** implemented a model-state manager enabling seconds-level rollback for harmful updates.
- **Outcomes.** Core techniques published as **OSDI ’22 (co-first author)**; subsequent production iterations enabled **10,000**× model-size scaling (GB → tens of TB per model) while maintaining **2.4s** model-update latency; deployed in WeChat recommendation stacks and **serves over one billion users daily**. After WeChat Channels fully adopted Ekko-based online recommendation, a WeChat official blog reports **+40%** DAU and **+87%** total VV over six months (alongside product iteration and operations).

Data engineering / feature platform: safe, scalable pipelines

- **Problem.** Modern feature pipelines are long and increasingly multimodal; cross-process operator composition creates high overhead and expensive data movement.

- **Approach.** Designed a WebAssembly-based runtime for **in-process isolation** (safety + resource constraints) and locality-aware operator placement near data sources.
- **Outcome.** Reduced data movement by up to **1,200**× on representative workloads; widely used within WeChat for data preparation.

LLM serving systems

- Building cost-effective serving mechanisms around remote KV-cache storage and compression.

Open Source

LLVM 2018 – Present

Developer (commit access)*; Google Summer of Code 2018 Participant*

- Improved Semi-NCA performance and optimization pipeline; shipped in LLVM 9.0 (reported speedups up to 1980× on real-world samples).
- Unified APIs on dominator trees; shipped in LLVM 7.0.

Academic Service

- **Reviewer:** CVPR 2025; ICLR 2025 Workshop on FM-Wild; NeurIPS 2025 Workshop on Efficient Reasoning.

Talks

- *Ekko: A Large-Scale Deep Learning Recommender System with Low-Latency Model Update.*
  Tencent WeChat AI Department (Shenzhen, Jun 2022); DataFun (Virtual, Aug 2022); TechBeat (Virtual, Sep 2022).

Selected Awards

- **Tencent Technology Breakthrough Award (Gold Prize)**, 2022H2 — **Project Lead, Ekko** (internal highest technical honor).
- Bronze Medal, **ACM-ICPC Asia Xi'an Regional Contest** (2017).
- Second Prize, **15th China Collegiate Programming Contest** (Guangdong Division) (out of 177 teams).

Selected Company/Press Write-ups

All links below are external write-ups about Ekko (OSDI '22).

- **WeChat official blog:** WeChat < 2.4 s
- **Tencent official blog:** "2.4 Seconds: Fast Enough!"
- **Synced Review / JIQIZHIXIN:** coverage