Chijun Sima

simachijun@gmail.com | LinkedIn: chijun-sima | Homepage: chijunsima.com

Education

South China University of Technology Guangzhou, Guangdong

B.Eng in Computer Science And Technology (Innovation class); GPA: 3.85/4.00 (Rank: 1/29) *September 2016 - June 2020*

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
  In 16th USENIX Symposium on Operating Systems Design and Implementation (OSDI '22), July 2022
- **Dynamic Barycenter Averaging Kernel in RBF Networks for Time Series Classification**:
  Kejian Shi, Hongyang Qin, **Chijun Sima**, Sen Li, Lifeng Shen, Qianli Ma
  In IEEE Access, April 2019

Professional Experience

Tencent

Senior SDE *July 2020 - Present*

- **MLSys**: Designed and implemented Ekko (OSDI' 22), a novel DLRS that enables low-latency model updates. It serves over a billion users daily and significantly reduces the model update dissemination latency compared to state-of-the-art systems. It is widely used inside Tencent and improves the profitability of several recommenders.
- **Isolation**: Designed and implemented Lightning, an SFI runtime powered by WebAssembly. It is later used for implementing storage functions inside Tencent's multi-tenant storage services. It is also in wide use as a serverless runtime inside Tencent. Compared with previous solutions (e.g., Splinter, OSDI '18; Shredder, Socc '19), it has tighter resource control on tenants and is more scalable by leveraging tiered compilation and compilation cache techniques.
- **Scheduling**: Designed Ekko Elastic by incorporating a model shard placement scheduler inside Ekko. It provisions machines by using operation research techniques on resource requirements of different ML models. It helps cut the cluster cost by up to 30% compared to static provisioning.
- **Distributed Systems**: Implemented a consensus algorithm library with WAN optimizations (e.g., parallel phase 2). Notably, it uses an adaptive batching algorithm, improving latency and throughput. It was later used for geo-replication at Tencent, powering the message queue system inside Wechat and serving over a billion users.

LLVM

Developer / GSoC 2018 Participant *May 2018 - April 2019*

- Improved the performance of the Semi-NCA algorithm implementation and worked to preserve the dominator tree along the optimization pipeline. Relevant changes were committed to LLVM 9.0.
- Unified API for updating Dominators in LLVM. The new API was committed to LLVM 7.0 and is in wide use.

Invited Talks

- **Ekko: A Large-Scale Deep Learning Recommender System with Low-Latency Model Update**:
  - Tencent Wechat AI Department, Shenzhen, June 2022
  - DataFun, Virtual, August 2022
  - Techbeat, Virtual, (expected) September 2022
  - South China University of Technology, Guangzhou, (expected) September 2022