

Tianyu(Terry) Sun

CONTACT <https://tianyu-sun.github.io>
INFORMATION <https://www.linkedin.com/in/tianyu-sun>

mobile: +1 (858) 214-0007
e-mail: tterrystsun@gmail.com

PROFESSIONAL
EXPERIENCE **Anyscale**
Software Engineer

Oct. 2023 – Present

- **[Anyscale Endpoints]** Orchestrated the design and execution of an embedding engine, streamlined its capabilities for deployment and serving. 7x24 service available at anyscale.com/endpoints.
- **[vLLM]** Optimized vLLM scheduler and reduced inter-token latency by up to 20%.
- **[vLLM]** Engineered a batched rotary embedding CUDA kernel, enhancing the efficiency of deploying multiple fine-tuned models. Also integrated fine-tuning support for Mixtral 8x7B model.
- **[Ray Core]** Enhanced the compiler interface of Ray compiled Directed Acyclic Graphs (DAG), which significantly reduces the control plane overhead.

SambaNova Systems
Senior Software Engineer

Apr. 2021 – Oct. 2023

- **[Graph Compiler]** Designed and implemented highly-scalable infrastructure for large-scale data and model parallelism, supporting multi-dimensional data parallelism, fine-grained hierarchical data distribution management, and efficient cross-socket traffic planning.
- **[Graph Compiler]** Designed and implemented compiler infrastructure for data parallelism on heterogeneous hardware, including bit-file packing and consistency checking support. US patents *US-20230237012-A1* and *US-20230237013-A1* granted.
- **[ML Performance]** Improved ML model resource and performance modeling at compile time and extended compiler resource modeling for multiple hardware architectures.
- **[ML Performance]** Brought up a compilation config development toolkit, resulting in a 10x increase in efficiency and widely adopted by hundreds of ML applications.

Aibee US
Research Intern

June 2020 – Sept. 2020

- Designed and implemented a model that improves the vehicle Re-ID performance by considering pose. Increased performance from 85.4% to 97.3% on TPR@FPR=0.01. Converted the PyTorch model to a Caffe model and shipped it to intelligent parking lot production.

Tencent
Tencent AI Lab

Dec. 2018 – Aug. 2019

- Participated in Virtual Host project, which aims at generating a virtual host for game streaming and weather broadcasting. Developed face segmentation and alignment modules, which were subsequently adopted by a million-DAU mobile application.

EDUCATION **University of California, San Diego**
M.S., Computer Science

Sept. 2019 – Mar. 2021

University of Science and Technology Beijing
B.E., Computer Science

Aug. 2015 – June 2019

SKILLS **Frameworks and Tools**

MLIR, PyTorch, Ray, vLLM, Kafka, TensorFlow, OpenCV

Programming Languages

C++, Python, CUDA, Golang, C

SELECTED
PROJECTS **Lego-Serverless Distributed Platform**

- Developed an event handling and function creation platform for modern serverless services.
- Designed a two-level load balancing mechanism consisting of a high-level round-robin load balancer and a middle-level Raft load balancer. Implemented data pipeline and high-level load balancing. Designed and developed data infrastructure based on Kafka and CouchDB.
- Lego-Serverless provides RESTful API for function and event CRUD. Additional management functions like user authentication and function authorization are supported too. The platform can handle 2,000 QPS in a single-node testing on an AWS EC2 instance.

Frame-GAN Gait Recognition Algorithm

- Proposed a method of boosting the accuracy of gait recognition by increasing the frame rate with generative adversarial networks. Published on *Neurocomputing* as *Frame-GAN*.