TIANYI CAI

Phone (Wechat): (+86) 18031521995 E-mail: tianyicai@vip.126.com

EDUCATION

Yale University *2015 - 2016*
M.S. in Computer Science, GPA: 4.0 / 4.0 *New Haven, CT*

University of Arizona *2011 - 2015*
B.S. in Computer Science, GPA: 4.0 / 4.0 *Tucson, AZ*

EXPERIENCE

Moore Threads 2025/02 - Present
LLM Training Platform Development *Beijing*

- Learning the operational mechanisms of large language model training frameworks based on Kubernetes, DL-Rover and Megatron-LM.
- Participating in developing a checkpoint storage system based on distributed memory and RDMA communication to improve single-node training recovery speed.
- Understanding the technical principles and product features of current advanced large models such as DeepSeek and Llama.

ByteDance 2021/10 - 2023/06
Douyin Recommandation Data Pipeline *Beijing*

- Maintained Douyin nearline data streams for video index and candidate databases. Enabled RocksDB based KV store to do column based updates and row based reads. Improved telemetry under multi-tenancy scenarios on the pipeline, avoiding particular user to overwhelm shared data buses.
- Optimized Douyin and TikTok video database storage by normalizing one monolithic table into video index and user index, improving user info consistency and cache efficacy, and simplifying user info lookup.
- Independently proposed, designed and implemented a Python light-weight streaming processing engine, and successfully deployed for vertical channels in Douyin such as local retail. Main features includes an *epoll* based event-driven loop for high throughput data transfer between Python processes, a multi-tier health monitoring and failure recovery mechanism on top of Kubernetes, seamless backward compatibility to minimize migration costs. By decoupling between vertical channels and Douyin main feed, they enjoy fast iteration and high reliability at the same time.

High-Flyer Quant 2021/03 - 2021/08
AI Infra (Firefly Super Cluster), Storage and Network *Hangzhou*

- Participated development of BeeGFS based parallel file system, improving write throughput by using Linux Asyc IO. Developed automated end-to-end testing on GitLab. Developed performance tests with visualization based on InfluxDB and Grafana.
- Designed and implemented asynchronous network library on RDMA networks. The library provides Boost ASIO like interface, event-driven communication pattern based on *epoll*. Both native Linux and Windows versions are provided, to support Windows apps for fast access to Linux storage cluster.

ByteDance 2019/08 - 2020/07
Web-Scale Search, Online Info Retrieval *Beijing*

- Maintained online services, including retrieval, fanout management, rank, request control. Improved dependency management for downstream services via unified registration, access and monitoring. Reduced search latency via mutiple tiers of caches. Improved content safety control quality by deploying streaming update on search indices.
- Restructured metadata storage in Xapian (search index library) by aggregating attributes into one protobuf structure, thus reducing number of metadata lookups during search retrieval phase, improved retrieval latency by roughly 10 percent.
- Responsible for architecture migration for Elastic Search to C++/Xapian/Hadoop based inhouse search infrastructure for various vertical channels. Lead initial efferts in designing strategies to enable inhouse search platform to support large number of small indices.

Google 2018/03 - 2019/06
GCP Service Mesh, Global Software Load Balancing *Sunnyvale, CA*

- Participated development of GCP Traffic Director product (Google managed Istio, now under the name GCP service-mesh management, acting as control plane to provide service-mesh related configs to Envoy proxies on data plane.) Implemented features including Envoy config cache, Envoy location awareness, Ingress authentication and authorizaiton configurations, etc.
- Developed deployment strategy for Envoy in hybrid cloud scenario. Environments includes VM, Kubernetes, GKE, either in customers' datacenters or Google Cloud.
- Developed integration test framework, which reads simple yaml config and automatically creates test workflows by calling Google Cloud API, substantially accelerated product iterations.

Datera, Inc 2016/09 - 2018/02
Scale-out Block Storage Cluster, Data Plane *Sunnyvale, CA*

- Participated development of data IO on top of Linux B-cache and Linux iSCSI. Maintained cluster-wise data operations such as consistency check and replica repair, storage load balancing among different storage nodes, etc.
- Helped in design and initial implementation of various features including encryption, de-duplication, and remote backup on AWS.

TECHNICAL STRENGTHS

Programming Languages	C/C++, Java, Python
Big Data & Distributed Systems	Hadoop, Apache Flink