TIANYI CAI

☎ (+86) 18031521995 ✉ 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

Cobo 2025/06 - 2025/08

MCP-Now, LLM Application Development *Beijing*

- Developed MCP-Now backend infrastructure, implementing PostgreSQL database access via Supabase Edge Functions. Built first version of user account system with secure authentication, registration workflows, and content sharing management. Established CI/CD pipeline using GitHub Actions and Supabase branch integration for automated testing and deployment.
- Leveraged AI development tools (Claude Code, Cursor) to enhance development efficiency and code quality.

Moore Threads 2025/02 - 2025/05

LLM Training Platform, GPU Ecosystem *Beijing*

- Developed LLM training platform integrating Kubernetes, DL-Rover, and Megatron-LM, providing one-click deployment, automated scaling, and real-time monitoring for large model training workflows.
- Developed high-performance distributed memory system leveraging shared memory, RDMA, and RoCE networking to provide checkpoint caching across training nodes, reducing recovery time from node failures.

ByteDance 2021/10 - 2023/06

Douyin Recommendation, Data Pipeline *Beijing*

- Developed and maintained Douyin's nearline data streams, video indices, and candidate databases. Developed RocksDB-based KV storage indexing module supporting column-based updates and row-based reads.
- Implemented comprehensive resource monitoring and governance mechanisms for multi-tenant scenarios, preventing individual services from overloading the main recommendation data pipeline and avoiding excessive index storage consumption.
- Improved Douyin and TikTok video database architecture by splitting monolithic index into separate video and user indices. Improved online cache hit rates and enhanced user information consistency while simplifying social feature data access.
- Independently designed and implemented a lightweight Python-based stream processing engine, successfully deployed for Douyin vertical channels including group buying. Core capabilities: *epoll*-based event-driven architecture for high-throughput IPC; Kubernetes-native health monitoring and automated failure recovery; full backward compatibility while facilitating business development iteration. Achieved faster feature iteration and reliability by completely decoupling vertical businesses from the main recommendation feed.

High-Flyer Quant 2021/03 - 2021/08

AI Training Platform (Firefly Super Cluster), Storage and Network *Hangzhou*

- Developed BeeGFS-based parallel file system with optimized asynchronous write performance via Linux Async I/O to improve read/write throughput. Built automated testing environment using GitLab CI/CD for VM auto-deployment and end-to-end testing, with performance visualization through InfluxDB and Grafana.
- Architected and implemented high-performance asynchronous communication library for RDMA over InfiniBand. Provided Boost.Asio-compatible API with event-driven architecture based on *epoll* for optimal performance. Developed cross-platform Linux and Windows native versions, enabling seamless integration between Windows applications and Linux storage clusters.

ByteDance 2019/08 - 2020/07

Web Search, Online Architecture *Beijing*

- Maintained and optimized Toutiao Search online services, providing web-scale search and internal search platform capabilities. Covered core modules including retrieval, fanout control, ranking, request control, and content safety.
- Managed downstream service dependencies through unified registration, access, and monitoring. Implemented multi-tier caching to optimize long-tail service responses, significantly reducing search latency.
- Implemented streaming updates for sensitive content flagging, enabling early-stage risk content filtering during retrieval, improving search quality while ensuring content safety.
- Restructured metadata storage in Xapian index library by aggregating document attributes into a single Protobuf structure, reducing metadata lookup operations during retrieval phase and decreasing retrieval latency by approximately 10%.
- Led architecture migration from Elasticsearch to in-house C++/Xapian/Hadoop infrastructure for multiple vertical search channels including user search and music search. Designed and implemented platform architecture supporting large numbers of heterogeneous small indices.

Google 2018/03 - 2019/06

Google Cloud Service Mesh, Global Software Load Balancing (GSLB) *Sunnyvale, CA*

- Participated in developing GCP Traffic Director product (managed Istio Pilot). Traffic Director serves as the control layer providing service mesh configurations to Envoy proxies on the data plane, enabling service discovery, load balancing, and authentication.
- Designed and developed core components including Envoy configuration cache, location awareness, and Ingress security configurations.
- Designed Envoy proxy bootstrap and linking methods across different environments including VMs, GKE, and native Kubernetes. Designed Envoy deployment solutions for both GCP and customer datacenters.
- Developed a Python-based integration test framework that automatically creates test environments by calling Google Cloud APIs based on configuration, significantly reducing manual operations and improving iteration efficiency.

Datera, Inc 2016/09 - 2018/02

Distributed Block Storage, Data Plane *Sunnyvale, CA*

- Participated in developing a distributed block storage system based on Linux B-Cache and iSCSI protocol, providing multi-replica read/write for storage volumes with support for automatic horizontal scaling and SSD/HDD hybrid configurations.

- Maintained core data plane functionality including inter-replica consistency check and repair, storage load balancing across different storage nodes.
- Contributed to the design of various product features including data encryption, deduplication across volumes, and remote backup from local datacenter to AWS.

TECHNICAL STRENGTHS

Programming Languages	C/C++, Python, Java
Systems	Distributed Storage, Stream Processing, Service Mesh, LLM Training
Databases	RocksDB, PostgreSQL, Key-Value Stores
Infrastructure	Kubernetes, RDMA/InfiniBand, Linux Kernel, CI/CD

Generated with the help from Claude.

source code: https://github.com/tianyicaii/resume