Yuelun Yu

Shanghai, China | +852 9488 2685 | ar8327k@gmail.com
github.com/ar8327 | LinkedIn/Yuelun Yu
Visa Sponsorship Required | CEFR C1, IELTS 7.5 | JLPT N2, Vocabulary: A, Grammar: A

Summary

Senior Software Engineer with 4 years of experience in building scalable backend infrastructure, APM systems, and ITSM platforms at ByteDance. Most recently led the development of an LLM Gateway supporting multi-model failover and unified API abstraction. Adept in Go, Python, and distributed systems. Passionate about performance optimization and system architecture.

Technical Skills

- Languages: Go, Python, Shell, SQL, Java
- Frameworks: SpringBoot, MyBatis, Gin, gRPC
- Databases: MySQL, Redis, Elasticsearch
- DevOps/Tools: Docker, Git, Kafka, RocketMQ
- AI/LLM: Multi-provider LLM orchestration, unified API abstraction, traffic shaping, high-availability design

Professional Experience

Senior Software Engineer - LLM Gateway Architect & Developer Feb 2025 – Present
ByteDance, Hangzhou, China

- Independently designed and developed multi-provider LLM Gateway supporting AWS Claude, Azure GPT, GCP Gemini and other vendors, providing OpenAI-compatible interface, deployed within weeks.
- Implemented high-availability load balancing system with session ID hashing to optimize Prompt Cache hit rate to 90%, achieved atomic queuing and rate limiting via Redis ZSET + Lua scripts, supporting 4-5k QPM.
- Established intelligent disaster recovery mechanism improving system availability from 70-80% to 99.9%; maintained 97% availability during GCP large-scale outages with automatic traffic switching.
- Developed tens of thousands of lines of provider adapter code for unified handling of different vendors' API formats, streaming responses, function calls, and error codes, with dynamic QPM adjustment and priority queuing.
- Built comprehensive monitoring system, discovered and helped GCP fix Gemini billing bugs through data analysis, securing significant refunds for the company, demonstrating exceptional problem identification and resolution capabilities.

Senior Software Engineer - APM Metadata System Architecture Design & ImplementationApr 2024 – Pre
ByteDance, Hangzhou, China

- Independently designed and implemented metadata management system supporting 2M services with 3000+ QPS and 99.9%+ SLA

- Implemented SQL-like DSL query engine with automatic field dependency parsing, unifying access to 20+ external data sources
- Designed multi-tier caching architecture (Redis + MongoDB) and intelligent degradation strategy, achieving zero impact from external service failures
- Built message queue-free task scheduling system using HTTP and distributed locks, processing 2M daily data updates

Senior Software Engineer - CMDB Platform Architecture Refactor Nov 2022 – Apr 2024
ByteDance, Hangzhou, China

- Core Achievement: Led the refactoring of million-scale CI CMDB platform serving 500+ internal systems with 5000 QPS daily queries.
- Refactored Python/Django services to Go microservices architecture, boosting single-instance throughput from 500 to 3000 QPS (6x improvement); eliminated Django ORM N+1 query problems and introduced multi-level caching, optimizing P99 latency from 10s to 200ms (50x improvement).
- Built MySQL + ElasticSearch dual-engine architecture, with MySQL handling transactional writes and simple queries, ES maintaining 100+ field wide tables for complex relational queries; designed Canal-based near-real-time data sync pipeline with RocketMQ message queue ensuring eventual consistency within seconds.
- Designed metadata-driven RESTful API framework dynamically handling CRUD operations for 30+ CI types, automatically parsing object relationships and converting to SQL; implemented row-level (based on data classification L2/L3/L4) and column-level fine-grained access control ensuring data security.
- Developed SQL-like DSL query language abstracting underlying MySQL/ES storage differences; intelligent routing distributes queries to optimal storage engines, supporting Scroll API to prevent deep pagination, limiting query depth to 1000 entries for system stability.
- Built stateless Worker clusters for data synchronization with message partitioning ensuring record ordering; implemented full data validation fallback mechanism with daily automated inconsistency repair; adopted Nginx weighted configuration for gray releases, ensuring zero-incident deployment over 3-month refactor project.

Software Engineer - IT Asset Management System Architecture Nov 2020 – Nov 2022
ByteDance, Beijing, China

- Designed and implemented server spare parts management system for 1M+ servers, covering 10+ component types (CPU, memory, HDD), supporting hundreds of daily server repairs with 30% reduction in end-to-end repair time.
- Built Redis-based distributed inventory reservation mechanism using atomic operations and two-phase confirmation to prevent overselling during concurrent part requests by multiple operators, ensuring eventual consistency.
- Designed multi-layer hardware compatibility rule engine with two-stage filtering (coarse screening + precise validation) supporting complex compatibility checks (CPU-memory joint validation, upward-compatible replacements).
- Applied Domain-Driven Design (DDD) principles, defining 4 core domain events (ticket creation, part inventory movements, repair completion) with async compensation via message queues and 2-hour retry window for fault tolerance.

Research Experience

Image Data Extraction Research — Nov 2019 – Mar 2020
Hangzhou Dianzi University

– Developed K-Means based segmentation and CRNN OCR pipeline for extracting data from scientific charts.

Short Video Content Filtering — Nov 2018 – Apr 2019
Hangzhou Dianzi University

– Built TextCNN-based classifier using TensorFlow for identifying illegal short texts, achieving 98% recall.

Publications & Patents

- Method for Automatic Extraction of Data in Images. CN201910972334.8, co-authored.
- Identifying Noisy Illegal Short Texts using Dual-Channel CNN. CN109670041A, co-authored.

Awards

- Second Prize, Service Outsourcing Innovation Competition, Ministry of Education, 2020
- Second Prize, Intel Embedded System Contest, 2018
- National Student Scholarship (2018, 2020), Ministry of Education
- School Merit Scholarship (2018, 2020), HDU

Education

Hangzhou Dianzi University — Sep 2017 – Jun 2021
B.Sc. in Information Systems, GPA: 4.18/5.0 — Hangzhou, China

Leadership & Activities

Vice President, MOOC Club — Jun 2019 – Jun 2020
HDU, Computer Science College

– Organized tech competitions and guided team formation for students.
– Delivered workshops and built mentorship system for younger peers.