

Fuchun Li	Target role: Senior Consultant Specialist (AI Gateway) 15 yrs hands-on engineering · 7 yrs architecture leadership · Tencent Cloud Certified Architect / Hall of Fame		
	Born Jan 1988 Gender Male Hometown Changning, Hunan Expected pay 30-45K RMB/mo Education Beihua University · Software Engineering · B.S. (Jul 2011) English CET-4 515 · professional reading/writing Phone/WeChat 13682320515 Email 505847426@qq.com Location Shenzhen (open to Guangzhou / Xi'an)		
Role Fit	1. AI / API gateway platform delivery: Built and operated LLM platforms with dual API gateway isolation (AgentX / DigitalHuman); multi-model LLM routing, tool calling, and RAG, covering design, development, test, deploy, and day-2 maintenance. 2. SRE & incident response: Hands-on monitoring, alerting, triage, and recovery (Prometheus / ELK / CAT / Zabbix); root-cause fixes and preventive automation under 99.9%+ SLO targets. 3. Gateway patterns: Multi-tenant isolation, authentication/authorization, rate-limit-style traffic control (Sentinel), and observability on open APIs; ready to deepen Kong / OpenResty plugin work on the job. 4. Enterprise Kubernetes: Production Kubernetes / Docker / Istio / GitOps across PaaS and SaaS; ingress and service-networking experience via Nginx / mesh. 5. Languages & async systems: Strong Java / Python for distributed services; productive with Golang for platform tooling; deep RabbitMQ and Kafka-class messaging for high-throughput async paths.		
Skills Fit	1. AI Gateway adjacency: Multi-provider LLM integration, prompt/tool pipelines, RAG, eval-minded iteration, and gateway-isolated service edges; ready to ramp on Portkey / APISIX / Higress. 2. Platform programming: Java / Python concurrency and performance in distributed systems; Golang for utilities/automation; Nginx-fronted gateways in production (Zenjoy). 3. Kubernetes in enterprise: Workload packaging, promotion, and day-2 ops for stateful/stateless services; cloud production ops (Tencent Cloud + private-cloud bank deliveries). 4. Messaging & throughput: RabbitMQ / RocketMQ / NATS-style async systems; daily push volumes from millions to ~10M with reliability gates. 5. Production support: User-facing troubleshooting, runbook discipline, and Tech Lead mentoring across 15 years of software engineering (5+ years platform/SaaS scale).		
Experience	Jul 2025 – Jul 2026 Microsoft Xiaoice (AI digital human / Agent) Platform Lead / Enterprise Delivery		
	1. Shipped DigitalHuman and AgentX on Docker/Kubernetes with dual API gateway isolation; cut end-to-end latency 2s → 500ms and held 99.9% availability for bank/insurance private clouds. 2. Owned multi-model LLM routing via MCP / Tool Calling, RAG pipelines, and day-2 ops (logs/metrics/alerts) for platform reliability. 3. Led Industrial Bank / Cathay / Ruizhong private-cloud deliveries with change-control discipline and production support under regulated workloads.		
	Nov 2024 – Jun 2025 ZTOGO / Gofu Express (cross-border last-mile) Backend Lead		
	1. Led microservice delivery for last-mile logistics; scaled capacity from 100K → 1M daily parcels across five countries. 2. Built high-throughput messaging (~10M pushes/day, 99.99% reach) and Flink + StarRocks visibility — SRE-style monitoring for delay and backlog risk. 3. Established design review, code review, and on-call runbooks for cross-module incidents.		
	May 2023 – Nov 2024 Keno (HeroDash global CX SaaS) Architect		
	1. Architected HeroDash on Spring Cloud + Kubernetes + GitOps; multi-tenant SaaS with open APIs, supporting 600+ concurrent sessions and 5M+ users. 2. Integrated Nacos/Sentinel/Seata for policy-style traffic control and resilience; elastic capacity cut ops cost 50%+. 3. Hardened bi-weekly release trains with observability for cluster health, queues, and API SLOs — production support for hybrid-cloud workloads.		
	Nov 2021 – Apr 2023 Zenjoy (game publishing) Java Architect		
	1. Rebuilt high-traffic publishing platform (~1.4B users, peak ~30K QPS); containerized CI/CD on Tencent Cloud with minute-level rollouts. 2. Deployed full-stack observability (ELK / Prometheus / Zabbix); seconds to detect, minutes to remediate production incidents. 3. Shipped Nginx-fronted async WebFlux gateway with circuit-breaking clusters for blast-radius control.		
	Apr 2020 – Nov 2021 Authine (Yunshu low-code PaaS) Architect		
	1. Built Yunshu's cloud-native stack from scratch: Docker + Kubernetes + Istio + Spring Cloud for HA and horizontal scale. 2. Stood up a 15-person CCB Beijing delivery team; three milestones to branch-scale rollout under bank change windows. 3. Raised automated deployment coverage to 90% and shortened delivery cycles ~40%.		
	Sep 2018 – Apr 2020 Lifesense (wearable health IoT) Senior Engineer		
	1. Owned multi-tenant health-cloud microservices and device ingest gateways; post-upgrade traffic hit 100M+/day at 99.99% availability. 2. Ran Docker/Kubernetes + Jenkins CD with CAT/ELK tracing for Linux production support. 3. Delivered Tianan Insurance health module; enterprise accounts grew to 300+.		
	Jan 2015 – Sep 2018 Qianhai Image (cross-border supply chain) Senior Java Engineer		
	1. Built Dubbo microservices for order/warehouse/logistics/finance from scratch with RabbitMQ-backed async flows; sustained ~1M orders/day. 2. Tuned end-to-end latency (order query 2.3s → 0.7s); introduced CI/CD and monitoring gates. 3. Drove code review and knowledge sharing; improved team delivery throughput ~40%.		
	Jul 2011 – Jan 2015 7Road (browser games) Java Engineer		
	1. Owned core platform modules (account/billing/CS/ops) under Linux production load. 2. Supported peak concurrency ~2,000 and ~200M lifetime users. 3. Maintained CMS/SEO sites for live-ops campaigns and brand conversion.		
Projects	AgentX Agent Engine LLM routing / tool calling Jul 2025 – Jul 2026 Context: Enterprise agent engine (Java + Python + Vue designer). Multi-source ingest, Agent orchestration, tool calling , and admin configuration — the control-plane work closest to an AI Gateway .		

AgentX 智能体引擎架构

接入层

API Gateway 双路隔离

Sa-Token 鉴权

AgentX API

Workflow API

拖拽设计器

编排层

Agent推理引擎(Java)

DAG workflows(Python)

多轮会话/记忆

任务规划/工具选择

能力层

多模型LLM

MCP工具适配

RAG/ES检索

业务工具

中间件层

MySQL

Redis

RabbitMQ

Elasticsearch

OSS/MinIO

基础设施

Docker/K8s

CI/CD

Prometheus/Grafana

Highlights:

1. Java **Agent** orchestration with multi-model **LLM** and **MCP tool calling** (sense → think → act).
2. **Python DAG Workflow** + designer: multi-step tasks, node scheduling, state, RabbitMQ async execution.
3. **Docker/Kubernetes** packaging with dual **API gateway** isolation; ship tool/rule updates without redeploying core business services.

Stack: Spring Boot + FastAPI + MyBatis-Plus + Sa-Token + Redis + RabbitMQ + ES + OSS + Celery + Docker + Kubernetes + LLM/MCP

DigitalHuman Platform | RAG / real-time LLM UX | Jul 2025 – Jul 2026

Context: Enterprise digital-human production and ops — LiveKit real-time, **RAG** Q&A, TTS/ASR, multi-engine streaming; bank/insurance private-cloud delivery.

数字员工平台架构图

客户端层



浏览器



小程序



Web SDK



移动端/Android/iOS SDK

接入层



防火墙/WAF



Ingress/OpenResty

前端容器层



员工平台前端(2C2G)



运营平台前端(2C2G)

应用层



员工平台后端(4C8G)



交互后端服务(4C8G)



LiveKit Server(4C8G)



运营平台后端(4C8G)



播报后台服务(4C8G)

算法服务层

播播视频生成服务集群



小样本视频推理(16C64G/7G GPU)



视频剪辑服务(16C64G)



LLM DeepSeek-R1
-Distill-Qwen-14B
(16C64G/24G GPU)



FunASR
语音识别
(4C8G/4G GPU)



LiveKit Agent
(4C8G/8G)



模型管理服务
(4C8G)



TTS集成中心
(4C8G)



TTS-Core服务
(32C128G)



TTS-WS服务
(4C8G)



资源中心服务
(4C8G)



FAQ服务
(Python)
(4C8G)

中间件层(数据&存储)



MySQL数据库



Redis缓存



Nacos配置中心



消息队列(RabbitMQ)



MinIO对象存储
(模型/音频/视频)



NAS/PVC 文件存储
(声音/形象模型等)



Elasticsearch(FAQ)

Highlights:

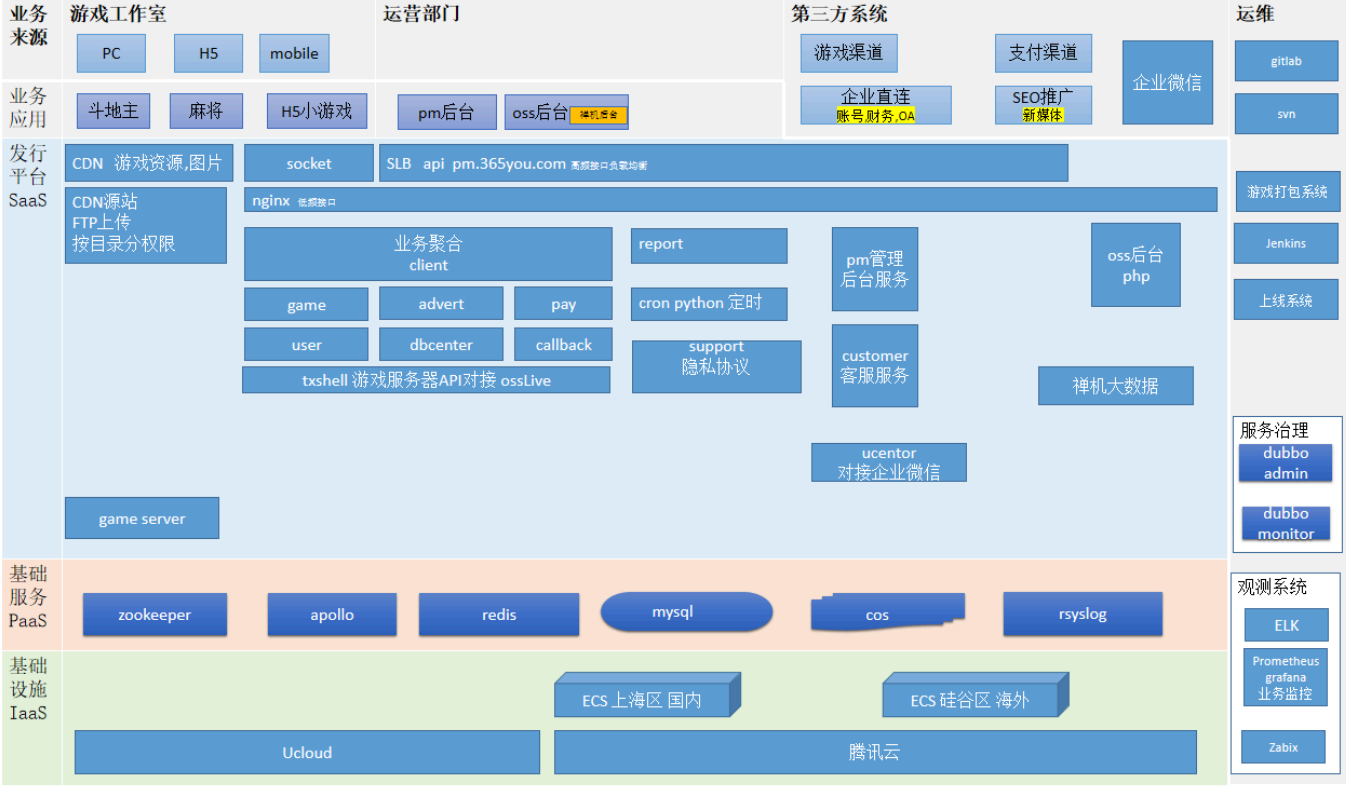
1. End-to-end ASR/TTS/LLM/RAG/LiveKit path; **99.9%** availability at ~10K-user scale.
2. Multi-engine TTS + SSE streaming; end-to-end latency 2s → **500ms** (75% reduction).
3. Private-cloud standard edition with on-prem middleware/DB adaptation for Industrial Bank / Cathay / Ruizhong.

Stack: Spring Boot + MySQL + Redis + RabbitMQ + ES + Nacos + MinIO + Docker + Kubernetes + LLM/ASR/TTS/LiveKit

HeroDash Global CX SaaS | Open API / multi-tenant | May 2023 – Nov 2024

Context: Hybrid-cloud contact-center SaaS (work.herodash.ai) — tickets, knowledge, AI assist, open APIs.

02 金箍棒部署架构 V2



- Highlights:**
1. ~1.4B users / ~9M DAU / peak ~30K QPS; Tencent Cloud containers + GitLab CI/CD.
 2. Nginx + WebFlux async gateway with circuit-breaking clusters.
 3. ELK / Prometheus / Zabbix: seconds to detect, minutes to clear incidents.

Stack: Dubbox + WebFlux + MySQL + Redis + ZooKeeper + Apollo + Nginx + Tencent Cloud containers

What I look for

1. Sustainable business with real production load
2. Healthy engineering culture and clear ownership
3. Modern AI platform / gateway / cloud-native stack

Tech advocacy / media



Tencent Cloud Toutiao WeChat CSDN GitHub

Targets