李鹏飞

sanfuyee@gmail.com

教育背景

中国科学技术大学	计算机应用技术	硕士	**2013.9~2016.6**
武汉大学	计算机科学与技术	本科	**2009.9~2013.6**

专业技能

1. 全栈架构与高并发实战：拥有 10 年后端研发及架构设计经验，深耕互联网金融（WeBank）、跨境电商（Shopee）及前沿 AI 工程（WPS）领域。具备从 0 到 1 搭建高可用基础架构的能力，实战经验涵盖 K8s、Istio、CI/CD 及 Prometheus 监控体系。
2. 多语言专家与分布式深度：精通 Golang、Java 及 Python 语言生态。深入理解微服务治理、DDD 领域驱动设计及分布式系统设计痛点；在复杂存储场景（TiDB 乐观事务、MySQL 优化）与大规模通信（Kafka、RabbitMQ、gRPC）方面有深度实践。
3. AI 工程化与大模型落地：掌握 NLP 理论，具备极稀缺的 MLOps 实战经验。主导过 DeepSeek-V3/R1、文生视频等大模型的 MaaS 服务架构；精通基于 vLLM、TensorRT-LLM 的推理加速优化及 RAG/Agent 框架落地，能有效降低算力成本并提升业务吞吐量。
4. 计算机底层与算法功底：具备扎实的计算机科学基础，深入理解数据结构、算法、网络及计算机组成原理；能够从底层原理出发解决系统性能瓶颈及复杂的工程难题。

工作经历

潞晨科技 **MLOps研发工程师** **2024.11~至今**

工作职责：

主导 MaaS (Model as a Service) 平台的架构设计与后端全栈研发。负责构建模型推理引擎与前端业务之间的核心中台，确保 AI 模型的高效、稳定与商业化落地。

项目经历：

- MaaS服务平台

  负责模多模型集成与交付， 成功上线自研 VideoOcean 文生视频、混元文生视频、Flux.dev 文生图及 DeepSeek-V3/R1 等主流大模型。商业化计费体系： 设计并实现了支持多种计费模式（如按量、按时、按次）的账单计费系统，并配套开发了模型市场与 API Key 管理模块。针对不同模型特性实现了精准的 TPM (Tokens Per Minute) / RPM (Requests Per Minute) 限流策略，保障多租户环境下的系统稳定性。开放并实现了三种不同维度的鉴权模式，在保障数据安全的同时满足了不同业务场景的接入需求。 从零搭建运营与监控告警面板，标准化梳理异常错误码体系，显著提升了系统的可观测性与运维响应效率。

技术栈：golang, go-zero, kubernetes, helm, kafka, redis, postgres

金山办公（WPS） **资深开发工程师** **2023.7~2024.9**

工作职责：

负责 WPS AI 大模型产品的落地与工程化。主导模型推理加速优化、后端业务系统开发及服务化部署架构，支撑高并发场景下的 AI 业务稳定运行。

项目经历：

- 模型服务

  基于 Python (FastAPI/Tornado) 构建高性能后端服务，支撑意图识别、阅读理解、PPT 自动大纲生成（RAG）及 PPT-Agent 等核心业务。深度应用 TensorRT-LLM 与 vLLM 框架，通过算子融合、权重量化、投机解码 (Speculative Decoding) 以及 Continuous Batching 等技术手段，显著提升了模型推理速度与系统整体吞吐量。基于 ReAct 框架 实现 PPT-Agent 自动化任务执行，整合计划（Plan）大模型、代理（Agent）大模型与召回服务，并实现了完善的历史会话管理机制。 在 PPT 大纲生成业务中，结合 BM25 算法与向量检索，利用 ElasticSearch 实现精准文档片段召回，实现大纲自动生成闭环。在意图识别业务中引入 Multi-LoRA 技术实现多场景共用 Base Model，大幅节省显存资源；在阅读理解业务中利用 vLLM 显存预分配 技术提升系统吞吐。

技术栈：python, fastapi, tornado, tensorrt-llm, tgi, vllm, tritonserver, continuous-batching, lookahead

粤港澳大湾区数字经济研究院（IDEA研究院） **AI全栈工程师** **2021.7~2023.6**

工作职责：

担任项目技术负责人（一号位），全面负责 AI 工程化方向。主导 AI 基础设施搭建、业务架构设计及核心模块开发；同时深度对接算法团队，支撑大语言模型（LLM）的训练、微调及推理全流程。负责团队建设，包括人才面

试、新人指导及任务拆解，直接向研究院领导汇报并落实项目需求 。

项目经历：

- GTSFactory模型自动化生产工厂

构建支持分类、实体识别、关系抽取及摘要生成的多场景平台，实现用户仅需少量数据即可自动训练适配业务的小模型闭环 。针对 Slurm 调度平台与云原生环境的差异，调研并集成 Kubeflow 机器学习平台 ；先后在华为云 ModelArts 与阿里云云原生 AI 套件上构建产品形态，利用网络隧道技术实现内网机器与阿里云 K8s 集群的深度融合 。负责 K8s 集群部署，支持 CPU/GPU 异构计算 ；基于 Elasticsearch 与 Milvus 搭建高性能向量引擎，助力训练过程中的数据增强 。建立完整的 CI/CD、监控告警（Prometheus/Grafana）及服务治理体系 。利用 FasterTransformers 进行推理加速，并基于 TritonServer 稳定交付 。实践 GPTQ、LLM.int8() 及 llama.cpp 等量化技术，优化模型在不同硬件环境下的部署性能 。

技术栈：

k8s, cicd, jenkins, prometheus, grafana, nacos, milvus, nlp, transformers, pytorch, gpt, tritonserver, fastapi, slurm, kubeflow, 模型量化，gptq, llm.int(), llama.cpp

深圳虾皮信息科技有限公司（Shopee） **高级开发工程师** **2018.10~2021.6**

工作职责：

作为骨干开发，对接产品经理，针对卖家端的各种业务需求进行需求分析，功能拆分，系统设计以及核心模块的开发。为提高研发效率，编写公共代码组件和开发规范，并指导培训新入职员工。基于Golang+Gin+GORM开发业务系统，充分使用和理解各种组件，例如Kafka, TiDB, Codis, gRPC等。

项目经历：

- Membership项目

会员项目给Shopee卖家提供了全套的会员方案，是卖家运营店铺粉丝的有效工具，提升了店铺会员的活跃度，增加了店铺的GMV转化率。整个项目在2个月内从0到1完成了设计和开发，我负责了其中的积分模块和权益模块，积分模块涉及订单过滤，订单转积分，入会送积分，积分过期等功能，权益模块涉及积分兑换商品，积分兑换优惠券，兑换重试等功能。其中设计和实现亮点有利用布隆过滤器过滤有效会员订单，调研并使用TiDB乐观事务模型，设计了合理的积分缓存方案，实现通用的接口幂等方案和业务重试系统，从API，存储，依赖，系统4个维度搭建了业务的监控。

- CRM系统

CRM系统本质是一个用于营销的Push系统，卖家在CRM平台过滤出特定的买家或者粉丝，并创建推送任务，向其推送文本、图片、商品或者优惠券信息，从而达到提高店铺GMV转化率的目的。我从0到1开发了CRM系统的核心模块，基于SOLID原则设计了推送系统的数据采集和数据推送流程，并完成了推送任务表从MySQL到TiDB的迁移。系统中多个服务之间使用gRPC通信，使用ETCD作为服务注册中心。CRM系统服务于60W卖家，大促期间每日推送用户数达到3千多万，达成了卖家提高GMV转化率的目标。

- WebChat系统

WebChat是一个服务于卖家和买家的即时聊天系统，我所在的团队专注在卖家PC端。卖家和买家间的消息会通过Kafka进行同步，最终使用WebSocket推送到客户端。项目主要基于Django框架开发，使用到Celery执行定时任务。消息会使用ElasticSearch保存，利用倒排索引实现了会话搜索。

技术栈：

python, django, celery, websocket, elasticSearch, go, gin, gorm, gRPC, protobuf, etcd, kafka, tidb, 微服务治理，分布式事务，solid设计原则，布隆过滤器，cap理论，raft协议，监控告警

微众银行（WeBank） **后端开发工程师** **2016.7~2018.9**

工作职责：

作为Java后端开发，承接部门内部各种零售业务的系统研发，包括微众银行APP登录平台、人脸审核系统、微众有折、钱包、礼品卡等。对产品经理的需求进行分析，并实现业务功能。与外部客户联调对接，共同达成业务目标。

项目经历：

- 微众银行APP登陆平台

登陆平台主要支持了微众银行APP和微众有折小程序的身份验证功能，基于OAuth2.0第三方授权机制，支持了微信登录、手Q登陆的功能。以微信登录为例，会首先引导用户微信授权，登录后台通过微信返回的code码，获取微信的accesstoken，并获取用户详细信息，再将微众银行内部账号和微信openID/unionID进行关联，实现了系统登陆态的转换。

技术栈：

java, springboot, springmvc, mybatis, mysql, redis, mumble-sdk, wemq, 两地三中心架构，网络分区