俞月伦

上海，中国 | +86 180 1707 8930 | ar8327k@gmail.com
GitHub/ar8327 | LinkedIn/俞月伦

英语水平：*CEFR C1*，雅思 *7.5* | 日语水平：*JLPT N2*，词汇参考级别：*A*，语法参考级别：*A*

个人简介

拥有四年后端开发经验的高级软件工程师，曾主导构建大规模分布式系统、应用性能监控（APM）平台和信息技术服务管理（ITSM）系统。近期负责字节跳动 AI 编辑器产品 Trae 的 LLM Gateway 设计与开发，具备多模型容错、高可用架构设计和统一 API 抽象能力。擅长 Go、Python 和分布式系统架构，热衷于性能优化与系统稳定性提升。

技术能力

- **编程语言：** Go、Python、Shell、SQL、Java
- **开发框架：** SpringBoot、MyBatis、Gin、gRPC
- **数据库与中间件：** MySQL、Redis、Elasticsearch
- **运维工具与消息队列：** Docker、Git、Kafka、RocketMQ
- **人工智能/大语言模型：** 多模型编排、统一 API 屏蔽、多云容错设计、限流与弹性调度

工作经历

高级软件工程师 - LLM Gateway 架构师 & 开发工程师 2025 年 2 月 – 至今

字节跳动，杭州

- 独立设计并开发多 provider LLM Gateway，支持 AWS Claude、Azure GPT、GCP Gemini 等多个服务商，提供 OpenAI 兼容接口，几周内完成上线。
- 实现高可用负载均衡系统，基于会话 ID 哈希优化 Prompt Cache 命中率至 90%，通过 Redis ZSET + Lua 脚本实现原子化排队和限流，支持 4-5k QPM。
- 建立智能容灾机制，将系统可用性从 70-80% 提升至 99.9%；在 GCP 大规模故障期间保持 97% 可用性，实现流量自动切换。
- 开发数万行 provider 适配代码，统一处理不同厂商的 API 格式、流式响应、函数调用和错误码，支持动态 QPM 调整和优先级排队。
- 建立全面监控体系，通过数据分析发现并协助 GCP 修复 Gemini 计费 bug，为公司争取到大额退款，展现卓越的问题发现和解决能力。

高级软件工程师 - APM 元数据系统架构设计与实现 2024 年 4 月 – 至今

字节跳动，杭州

- 独立设计并实现支撑 200 万服务的元数据管理系统，QPS 3000+，SLA > 99.9%

- 实现类 SQL 的 DSL 查询引擎，自动解析字段依赖关系，统一了 20+ 个外部数据源的访问
- 设计多级缓存架构（Redis + MongoDB）和智能降级策略，外部服务故障零影响
- 基于 HTTP 和分布式锁实现无消息队列依赖的任务调度系统，每日更新 200 万条数据

高级软件工程师 - CMDB 平台架构重构 2022 年 11 月 – 2024 年 4 月

字节跳动，杭州

- 核心成就：主导百万级配置项 CMDB 平台重构，服务 500+ 内部系统，日均承载 5000 QPS 查询请求。
- 将 Python/Django 服务重构为 Go 微服务架构，单实例吞吐量从 500 提升至 3000 QPS（6 倍提升）；通过消除 Django ORM 的 N+1 查询问题和引入多级缓存机制，将 P99 延迟从 10 秒优化至 200 毫秒（50 倍提升）。
- 构建 MySQL + ElasticSearch 双存储引擎架构，MySQL 处理事务性写入和简单查询，ES 维护 100+ 字段的宽表支撑复杂关联查询；设计基于 Canal 的准实时数据同步管道，通过 RocketMQ 消息队列确保数据最终一致性，延迟控制在秒级。
- 设计元数据驱动的 RESTful API 框架，动态处理 30+ 种 CI 类型的 CRUD 操作，自动解析对象关联关系并转换为 SQL；实现行级（基于数据密级 L2/L3/L4）和列级细粒度权限控制，保障数据安全。
- 开发类 SQL 语法的 DSL 查询语言，屏蔽底层 MySQL/ES 存储差异；通过智能路由将查询分发到最优存储引擎，支持 Scroll API 防止深分页，限制查询深度在 1000 条以内保证系统稳定性。
- 构建无状态 Worker 集群处理数据同步，通过消息分区保证同一记录的顺序性；实施全量数据校验兜底机制，每日自动修复数据不一致；采用 Nginx 权重配置实现灰度发布，确保 3 个月重构项目零故障上线。

软件工程师 - IT 资产管理系统架构 2020 年 11 月 – 2022 年 11 月

字节跳动，北京

- 设计并实现百万级服务器备件管理系统，覆盖 CPU、内存、硬盘等 10+ 种部件类型，支撑日均数百台服务器故障处理，端到端维修时长缩短 30%。
- 构建基于 Redis 的分布式预扣库存机制，通过原子操作 + 二次确认策略解决多运维人员并发申请部件的超卖问题，保证库存数据最终一致性。
- 设计多层硬件兼容性规则引擎，采用特征粗筛 + 精确校验的两阶段过滤模式，支持复杂兼容性判断（如 CPU-内存联合校验、向上兼容替换），提升部件匹配准确率。
- 实践领域驱动设计（DDD），定义 4 个核心领域事件（工单创建、部件出入库、维修完成），通过消息队列实现异步补偿与 2 小时重试窗口，确保系统容错性。

科研项目

图像数据提取算法研究 2019 年 11 月 – 2020 年 3 月

杭州电子科技大学

– 构建基于 K-Means 图像分割与 CRNN OCR 的图表信息抽取系统。

短视频文本内容过滤系统 2018 年 11 月 – 2019 年 4 月

杭州电子科技大学

– 基于 TextCNN 和 TensorFlow 实现非法短文本识别系统，召回率达 98%。

专利与论文

- 一种图像中数据自动提取的方法，发明专利，专利号：CN201910972334.8，合作作者
- 基于双通道卷积神经网络识别噪声非法文本的方法，专利号：CN109670041A，合作作者

荣誉奖项

- 教育部服务外包创新大赛全国二等奖（2020 年）
- 英特尔嵌入式系统设计大赛全国二等奖（2018 年）
- 国家奖学金（2018、2020 年）
- 杭州电子科技大学校级综合奖学金（2018、2020 年）

教育背景

杭州电子科技大学 2017 年 9 月 – 2021 年 6 月

信息系统本科，*GPA*：*4.18 / 5.0* 中国 · 杭州

社团经历

MOOC 编程俱乐部副主席 2019 年 6 月 – 2020 年 6 月

杭州电子科技大学计算机学院

– 组织编程比赛，指导成员组建项目小组参与实战。

– 开设工作坊并建立新生导师制度，提升学生参与度与项目质量。