

李孟凯 (Peter)

效率研发基建 · 底层分布式集群调优 · Agent 弹性计算

北京
limengkai@yuemake.work
https://l1.ft
github.com/mengkaibulusili

核心优势

- 大规模构建集群 · 深度调优：华为云云构建 CAS 800T，等效节约 38 万核 CPU / 760 TB 内存；独立运维单集群 上千节点；bazel 分布式编译集群（动态扩缩容 + Mnemonic 路由分组 + CAS/Worker 分离，构建资源 -70%）。
- 源码级底座与网络：K8s / OpenStack 源码开发；NFS CSI Driver 快照与克隆、分布式 CSI、极速快照、COW、dentry / dcache；QEMU 逐行 debug 内核启动，perf / strace 跨层定位。网络覆盖单机协议栈到交换机（iptables / nftables、WireGuard、tcpdump、eBPF、K8s Calico），自研流量可观测与流量镜像分析组件。
- Agent 弹性计算：Agent 沙箱快照与快速恢复、容器化沙箱、任务编排与并发限流（Cron 定时、批量推理、异步队列、失败重试）；自建 LLM 网关支撑 Claude Code / Desktop 等 Agent 客户端规模化接入。

有趣探索

- QEMU 逐行 debug 内核启动：单步跟踪引导链与内核初始化，定位启动失败的根因，并打通 debug 流量发送 链路。
- libc / libcpp 段错误定位：崩溃栈在标准库，根因却是上层软件的非法调用；回溯到真正的调用方，核对参数约定与对象生命周期，从业务侧规避而非改库。
- 私有云除零与硬件级故障：从 dmesg 区分软件除零与硬件异常，联合厂商复现故障、定位根因，最终升级 flash 固件解决。
- 为 NFS CSI Driver 增加快照与克隆：补齐 VolumeSnapshot 与卷克隆特性，接进 K8s 存储链路。
- 飞书外接 Claude SDK Agent 服务：连通研发业务数据读取与服务调用路径，24 小时高准确率定位疑难问题，为 500+ 车企研发大群提供 oncall 支持。
- 业余硬件：ESP32 手机蓝牙配网盒子、AI 语音盒子；北京 / 武汉 / 长沙三地 WireGuard 跨城组网，重编 ARM 内核启用 WireGuard 内核模块，单台成本约 50 元。

AI 消耗量（自建网关 new-api 模型统计）

975.20M token / 8h 单模型（deepseek-v4.1-flash-exp） 288.7 t/s 峰值吞吐 97.9% 缓存命中 ≈8.04B token / 近 7 天（网关可监控口径）

以上仅为自建网关可监控口径；另有多套未纳入网关监控的 Codex 账户在用，实际消耗更高。

工程理念

- 构建速度即迭代速度：编译与 CI 的每一秒都会被团队人数与每日提交次数放大——应将其做成有 SLA 门禁的工程指标。
- 车灯效应：车灯照射的距离决定能开多快——可观测的深度，决定团队能安全推进的速度上限。
- 大型 monorepo 治理：日均 5,000+ PR 的仓库，治理标准只能靠工具强制——单 Job 部署契约、增量构建缓存、统一制品源，都不能靠人自觉。

工作经历

某具身智能企业

2026.08 – 至今

AI 工程化平台 / 效率研发基建负责人

- 建设统一构建缓存与镜像体系、三级缓存与统一出口，test / prod 部署 80s / 110s；治理 CCE K8s 1.34 双集群与 10.30.0.0/16 VPC 七子网跨四可用区。
- 自建 LLM 网关（new-api 定制 fork）三协议 / 三渠道路由、tiered_expr 分时计价、Logto OIDC + Passkey；定位并修复 LLM 缓存失效，旧客户端命中率 0% → 99.7%。
- 建设可观测告警栈与数据采集平台；全部线上变更均由 IaC 定义，iac diff 保证 repo↔live 零差异。

轻舟智航

2025.04 – 2026.08

AI Agent DevOps 平台负责人

- 服务奇瑞、吉利、理想等多家头部车企：从零建设合规云与 CI + 构建系统，支撑每天 上千个仿真测试版本 的发布，获甲方高度认可并收到其 offer。
- 打通飞书外接 Claude SDK Agent 服务：连通研发业务数据读取与服务调用路径，24 小时高准确率定位疑难问题，为 500+ 车企研发大群提供 oncall 支持。
- 独立搭建 Agent 云平台并接入飞书群聊机器人，支持群 / 机器人级 Prompt Template、Skills/MCP、工具权限与任务模板；设计 Cron、批量推理、异步队列、失败重试、并发限流与容器化沙箱。
- 研发并调优大规模分布式编译集群，动态扩缩容与任务路由降低 70% 构建资源消耗；单人运维 上千台 构建集群并二次开发调度框架。

九维数智（北京）科技有限公司

2024.02 – 2025.04

云原生研发工程师

- 为 NFS CSI Driver 增加 快照（VolumeSnapshot）与卷克隆 特性，涉及 K8s CSI 驱动与 Operator 二次开发、分布式 CSI、极速快照、COW 与 dentry / dcache 路径。
- 建设 LDAP 成员管理与 Code Review 审核服务；构建 CI/CD 平台（Gerrit / Jenkins / BuildKitd / Bazel / Harbor）；带领 10+ 人团队推进研发流程标准化。

华为云计算技术有限公司

2022.03 – 2022.12

华为云 · 云龙 DevOps · 分布式构建系统研发

- 参与华为云云龙 DevOps 高性能分布式编译系统研发，服务 昇腾、鸿蒙 两个超大仓库源码构建；构建 CAS 缓存 800T，缓存命中 95%+ 时仍占 2 万核 / 40 TB，相当于节约 38 万核 CPU / 760 TB 内存。
- 定位重复文件下载与重复构建问题并增加 fastpath，缩短构建时间 70%，助力团队获得总裁嘉奖令；基于 Bazel、BuildBarn、ccache 落地分布式增量构建缓存，与欧洲团队协作参与 justbuild 开源项目。

武汉泽塔云科技股份有限公司

2021.09 – 2022.03

云计算研发工程师

- 以 QEMU 逐行 debug 定位内核启动问题（strace、perf、overlayfs 与容器根文件系统构建）；OpenStack 私有云平台二次开发与运维、SPICE 虚拟桌面研发。

自由职业

2023.01 – 2024.02

独立开发者 · 抖音 / 微信小程序

技术栈：Kubernetes 1.34 / CCE 双集群 · BuildKit 0.30.0 · Bazel / BuildBarn · Terraform · Nexus · JuiceFS · Prometheus / Grafana / Loki · iptables / nftables · WireGuard · tcpdump · eBPF · Calico · Go / Python / Shell / C

GitHub：github.com/mengkaibulusili —— sglang / libfuse / buildfarm / justbuild 与 Linux 内核源码的持续跟踪与实践