郑宇宸 Yu-Chen CHENG

https://rudeigerc.dev · rudeigerc@gmail.com

个人简介

郑宇宸是一名专注于大规模 AI/LLM 基础设施和 Kubernetes 云原生平台的软件工程师。拥有 2+ 年架构生产级 AI/LLM 系统的经验，他参与了和鲸科技的关键 LLM 基础设施项目，并作为 AI 平台开发工程师推动了宽德投资的机器学习平台现代化。他将分布式系统的技术专长与在跨职能协作和建立技术标准方面的能力相结合。他也积极贡献包括 Kubeflow 在内的 CNCF 项目。

教育经历

上海交通大学 中国 上海

工程硕士（专业学位），软件工程 *2019 年 09 月至 2023 年 03 月*

学位论文：**多租户共享集群中的分布式数据处理作业调度**

上海交通大学 中国 上海

工学学士，软件工程 *2015 年 09 月至 2019 年 06 月*

工作经历

和鲸科技 中国 上海

软件工程师 *2025 年 05 月至今*

- **LLM 后训练**：深度参与架构设计并主导实施了基于 Kubeflow Trainer 的大语言模型后训练流水线系统，其采用云原生架构，实现了高度可扩展的数据管理、分布式模型训练和多维度评估工作流，为用户提供服务的生产级 LLM 部署。
- **LLM 推理**：参与从 Knative 到 AIBrix 的模型推理技术迁移项目，重新架构设计了推理服务 API 体系，引入了基于前缀的智能路由算法和基于时间序列预测的水平自动扩缩容机制，在关键性能指标上实现了显著优化。同时实现了全面的 LLM 推理性能基准测试框架，以评估不同模型架构和部署配置下的性能指标。
- **机器学习平台**：与产品经理和算法研究团队深度协作，参与制定 AI/LLM 基础设施的技术需求规范和长期技术演进路线图。通过构建基于 GitLab Runner 的 CI/CD 流水线和集成 OpenTelemetry 实现的全链路分布式追踪系统，大幅提升了平台的稳定性、可观测性和故障排查效率。

Kubernetes Go Python LLM LLMOps

宽德投资 中国 上海

AI 平台开发工程师 *2023 年 05 月至 2024 年 09 月*[1]

- **AI 平台架构**：深度参与架构设计并主导了基于 Kubernetes 的内部 MLOps 平台的全面重构工程。推动了 Python SDK 和 CLI 工具链的现代化升级、基于 Volcano 与 KubeRay 设计实现了统一的 Job 抽象模型体系、基于 Protocol Buffers 和 gRPC 构建了高性能微服务通信框架，并开发了基于 React 和 Vue.js 的现代化全栈管理仪表板。
- **DevOps 流水线工程**：设计并从零构建了基于 Tekton Pipelines 的企业级 CI/CD 流水线系统，支持 Python 和 Go 项目的自动化构建、测试和部署。通过集成代码质量检测和自动化测试框架，将开发和测试周期缩短了 40%，大幅提升了团队开发效率和代码质量。
- **技术领导力与标准化**：a) 制定并推广了基于 PDM 的现代 Python 项目管理标准和最佳实践规范。b) 设计实现了用于在 Kubernetes 集群上自动化部署静态网站的 Helm Chart 模板，深度集成了 Tekton 流水线和 OAuth2 Proxy 认证中间件。c) 开发了基于 VitePress 的统一技术文档生成和发布平台，建立了完善的文档工程化体系。
- **技术指导**：承担导师角色，指导完成了 a) Dex 身份认证提供器与企业 LDAP 系统的集成开发，为内部微服务体系构建了统一的身份认证与访问控制中间件。b) Kubernetes 多调度器性能基准测试框架的设计与实现，提供了用于评估和对比不同调度算法性能的标准化测试平台。

Kubernetes Go Python Ray MLOps

蚂蚁集团 中国 上海

实习站点可靠性工程师 *2021 年 06 月至 2021 年 09 月*

- **联邦可观测性平台**：参与了内部 Kubernetes 多集群可观测性平台的初期设计与开发，实现了跨分布式 Kubernetes 环境的统一监控。

[1] 2024 年 10 月至 2025 年 02 月作为军事训练役在台湾服役。

- **身份提供者连接器**：主导设计并实现了连接企业内部身份与访问管理 (IAM) 系统的 Dex OIDC 身份提供器插件。该组件采用标准 OAuth 2.0 和 OpenID Connect 协议，为包括 Argo CD、Grafana 在内的云原生工具链提供了统一的单点登录 (SSO) 和细粒度权限控制服务，显著提升了企业内部系统的安全性和用户体验。

Kubernetes | Go | React

Linux 基金会 远程
LFX 导师计划（LFX Mentorship Program）学生， CNCF - Volcano *2021 年 02 月至 2021 年 05 月*

- **系统稳定性优化**：Volcano 是一个 Kubernetes 原生的批处理调度系统，用于计算密集型工作负载。搭建了高效的基于 Kubernetes 的测试框架，并且导入单元测试和 E2E 测试，以覆盖更多的使用场景。

Kubernetes | Go

服务

- Google Summer of Code 2025 导师， Kubeflow， Project 10: Support Volcano Scheduler in Kubeflow Trainer

编程技能

语言：Go, Python, TypeScript, C++
技能与专长：LLMOps, MLOps, Resource Management, Scheduling
工具与框架：Kubernetes, Kubeflow, Ray, React, Vue

证书与认证

日本语能力测试 JLPT N1 2017

其他

博客：https://rudeigerc.dev
GitHub：https://github.com/rudeigerc
LinkedIn：https://www.linkedin.com/in/rudeigerc/
Medium：https://medium.com/@rudeigerc
知乎： https://www.zhihu.com/people/rudeigerc
微信公众号：YC Cheng (@yuchenrcheng)
语言能力：中文（母语）、英语（专业工作能力）、日语（有限工作能力）

开源项目贡献

- 组织成员：Kubeflow, InftyAI, Tekton 和 Tensorchord
- 贡献：kubeflow/trainer, InftyAI/llmaz, kubernetes-sigs/inference-perf, tektoncd/triggers, volcano-sh/volcano, tensorchord/envd, ray-project/kuberay, pdm-project/pdm

Last Updated: 2025-09-03