

# 徐鑫

Infrastruct Tech Leader

- +8613603043209
- 277040271@qq.com
- www.dreamrounder.com
- 广州



我目前就职于字节智能创作-AI平台部门，该部门提供了抖音剪映(C端)以及火山引擎(B端)的绝大多数AI能力。而我则负责其中基础架构方向的研发团队，涉及的主要技术包括网关、模型存储、工程框架、SRE、研发规范以及推理交付平台等，上接AI工程团队，下接字节基础设施，为团队打造高质高效的AI交付链路。此前在腾讯游戏负责容器平台的研发，为上层业务提供推理与训练平台，推动大量游戏业务向云原生转型，积累了丰富的云原生经验。工作之余也热爱参与开源，为ApacheAPISIX 的 PMC Member，也贡献/参与过不少开源项目，详情请移步 Github个人页。

## WORK EXPERIENCE

### AI Infra Tech Leader

ByteDance | 2021 - Present

带领AIP的基础架构团队，从B端业务切入，积累了可靠的基础技术解决方案，并以此为基础，将野蛮生长的C端平台技术进行了技术革新，在工程交付、服务运维、技术基建等几大方向上，一边偿还技术债务一边从 0 到 1 去构建体系化的解决方案，在各个方向上都取得了超乎预期的成果

- 搭建交付中心、自动工程化、网关控制面、Mesh WASM并结合开源的ComfyUI方案，将交付效率从 > 5d 提升到零工程交付的程度，无需投入工程化人力，业务自行在平台自助进行交付，截止2024年为止平台交付的AI能力过千
- 通过镜像瘦身、分层，ModelMultiplexing，模型分级缓存，依赖分发加速，自动扩缩容，智能运维等手段把整体推理GPU的日均利用率从20%提升到了50%，部分方向AI能力在80%以上
- 从 0 到 1 构建资源中心解决方案，用于解耦业务对算力与存储资源的依赖，实现混合多云的资源调度能力。在计算资源上支持了音视频转码、编辑、模板创作等核心能力的多云调度与动态切流，达到分钟级别的异构容灾。在存储资源上支持了基于磁盘的分级缓存、数据加密、秒传、并行传输等核心能力，作为所有线上服务的底层对象存储，日峰值QPS10w+，峰值带宽100GB/s以上
- 孵化出包括网关、IAM、多云混合资源调度、私有化架构、推理框架与研发基础库等基础能力，覆盖日常研发50%以上的需求场景，极大提升研发效能，规范日常研发质量，支撑BC端70%以上的平台需求与AI工程场景
- 组建完善的SRE体系，涉及商品化、稳定性与资源管理。商品化方向上，建立了完整的定价模式与自动估价体系，让之前混乱的账单回归正常，并在2024年配合业务完成亿级别的成本梳理；稳定性上沉淀了BC端共用的治理规范（可观测性、容灾、应急预案等方向），并以此推动BC端各个方向完成了大量从一个9到四个9的转变；资源管理上通过拉起各类技术优化专项，完成GPU资源利用率从20%->50%的优化。

### Senior Software Engineer

Tencent | 2019 - 2021

主导平台的微服务架构设计和开发，同时为我们的用户提供技术咨询与集群资源管理，平台使用k8s提供调度能力，GO编写管控服务，截止2020年底，平台服务于公司数十种产品，部署POD数十万。

- 为平台搭建一套完整的分布式指标采集组件，负责采集整个平台百万级别 POD 的性能指标与业务指标，截止 2020 年底，每秒写入指标超过千万级别。
- 技术支持 10+ 项目从传统架构转型到云原生，其中包含日访问量亿级的项目
- 编写的高性能分布式任务流框架在中心内推广到了各个小组，帮助解决复杂的工作流，如k8s、redis集群的运维管理，富容器的管理等。
- 基于自研的工作流框架实现了 k8s 集群节点的自动伸缩与快上快下，只需指定期望数量，其他一切自动完成。
- 负责整个平台的基础设施，包括网关和鉴权等，保证平台的稳定，截止 2020 年，日请求量千万级别。
- 在公司内推广 ApacheApisix，支持 7 款产品成功落地网关

### Software Developer

万汇互联 | 2018 - 2019

负责社交产品的网关开发与微服务架构设计开发

- 引入 Kong 作为网关，解放了微服务在AOP层的工作
- 使用DSL构建了一个事件系统，优雅地解决了基于用户行为的虚拟币分发功能

### Software Developer

深圳巨鼎医疗设备有限公司 | 2016 - 2018

负责医院报告打印系统的基础框架设计和开发、带领 web 团队完成业务需求

- 在两年内通过不断沉淀、优化基础框架与组件，帮助公司核心产品愈发稳定，最后成功孵化，只留下定制化工程师，核心团队解散
- 第一次带团队，带领 3 人左右的小团队按时按质完成 web 需求，并且完成了从传统 jQuery -> MVC 的 web 转型

### Assistant Software Developer

AbeamSystem | 2014 - 2016

负责ERP系统的二次开发与文档编写

- 接触到了很多项目，所以语言栈丰富，包括: c#/VB/c++/js 等
- 由于日企的代码都需要反复review，培养了自己对代码质量要求高的习惯
- 在日本的一年里，不断提升日语技能，能够正常以日语和同事进行工作交流

## PROJECTS

<b>ApacheApisix</b>	PMC
Apache下毕业最快的顶级项目之一，是一个基于 Openresty 的网关项目，在腾讯服务于数个百万DAU的产品。作为其PMC，为其贡献过诸如 CORS, BatchRequest等特性，同时也会参与项目的设计评审与CR。	
<b>Kubernetes</b>	Developer
内部云平台的调度层，是我们为用户提供基础服务的根本，我一般会编写一些 controller、webhook 以及 operator 来为平台用户提供特性功能。	
<b>Prometheus</b>	Developer
CNCF 的毕业项目，云原生下的观测性大多都是基于本项目实现，我用于为整个平台提供指标监控和采集方案，采集每秒千万级别的性能指标和业务指标。	
<b>go-restful</b>	Contributor
k8s 用于内部组件通信的 http 框架，没有任何外部依赖，同时性能也较高，该框架也是 go-chassis 的路由框架，我为其贡献了GoogleAPI 设计规范的 CustomMethod.	
<b>go-chassis</b>	Contributor
华为开源的微服务框架，我用于内部微服务开发，迭代过程帮助项目提升了性能与稳定性，因此受邀作为讲师参加了华为赞助的Gopher线下沙龙。	
<b>hydra</b>	Contributor
一个实现了oauth2的开源项目，我用其实现了平台的Oauth2授权，在使用过程中为项目贡献了一些fixs。	

## EDUCATION

电子信息工程  
大连大学  
2010 - 2014

## AWARDS

<b>ApachePMC</b> ApacheApisixPMC 2020
<b>系统架构设计师</b> 软考高级职称——系统架构设计师 2019
<b>PMP</b> 项目管理专业人士资格认证 2019

## SKILLS

<b>语言栈</b> Go/DotNet/Javascript  Lua/C++/Python  HTLM/Css
<b>存储</b> Mysql/SqlServer/Oracle  Mongo/Etcd/Redis  RabbitMQ/Kafak/Es

## LANGUAGES

中文 (Native)
英文 (Good RW/Poor LS)
日语 (Good)

---

