才天一

电话（微信）：18031521995　邮箱：tianyicai@vip.126.com

教育背景

耶鲁大学 *2015 - 2016*

硕士，计算机科学，GPA: 4.0 / 4.0 *New Haven, CT*

亚利桑那大学 *2011 - 2015*

本科，计算机科学，GPA: 4.0 / 4.0 *Tucson, AZ*

工作经历

摩尔线程 2025/02 - 现在

大模型训练平台开发 北京

- 学习基于 Kubernetes, DL-Rover 和 Megatron-LM 的大语言模型训练框架的运行机制。
- 参与开发基于分布式内存和 RDMA 通讯的 checkpoint 存储系统，提升单节点训练恢复速度。
- 了解 DeepSeek, Llama 等当前先进大模型的简要技术原理和产品功能。

字节跳动 2021/10 - 2023/06

抖音推荐离线架构 北京

- 维护抖音视频近线数据流以及视频正排和候选库，使基于 RocksDB 的 KV 索引同时具有按行读取和按列进行流式更新的能力。完善在多业务场景下对数据流和索引存储复用时对资源占用的监控和治理，避免个别业务的高负载对主要推荐数据流造成积压以及索引存储空间的过度消耗。
- 对抖音和 TikTok 视频索引库按视频和作者两个维度进行拆分，抽取视频中的作者信息构建独立的用户索引，以减轻视频索引库的存储压力，提升在线系统缓存的命中率，增强作者信息的一致性，方便社交等业务对作者信息的获取。
- 独立设计并开发了一个基于 Python 的轻量化流处理引擎，完成在抖音团购等新业务中的落地。设计采用基于 *epoll* 的事件驱动实现多进程的高效异步通信，基于 Kubernetes 实现对节点的健康监控和错误恢复，通过配置信息可灵活更改吞吐和并发量，采用 Python 以实现对历史代码的无缝兼容以及方便算法同学的后续业务开发。新的数据流在提升新业务迭代速度的同时，通过和抖音主端解耦，极大提升了抖音主要数据流的稳定性。

幻方量化 2021/03 - 2021/08

机器学习训练平台（萤火集群）存储和网络开发 杭州

- 参与开发并行文件系统，对 BeeGFS 进行基于 Linux Async IO 异步写入的改造。开发自动化集成测试环境，基于 GitLab 在虚拟机自动部署和测试，并基于 InfluxDB, Grafana 实现测试结果的可视化。
- 设计并开发了 InfiniBand 网络下基于 RDMA 协议的异步网络通讯库。设计采用了类似于 Boost Asio 通讯库的 API, 基于 *epoll* 实现异步通讯, 并且分别开发了 Linux 和 Windows 的原生版本, 以支持 Windows 业务应用对 Linux 存储集群的高效访问。

字节跳动 2019/08 - 2020/07

头条搜索在线架构 北京

- 维护并优化头条搜索在线服务，涉及召回模块、扇出控制、排序逻辑、请求控制、内容安全和特殊产品需求。治理在线服务下游依赖，规范依赖服务的注册和调用，监控下游请求延迟，通过添加缓存以优化部分长尾下游的耗时。对敏感内容标记做流式更新改造，使其能更及时且更早地在检索流程中被消除，保证内容安全的基础上提升搜索效果。
- 优化倒排索引 (Xapian) 对文章 metadata 的存储和读取方式，对一个文章的 Meta 信息做聚合，在召回初段做统一读取，以减少 Xapian 多次对同一文章的不同 meta 信息的搜索，缩减召回延迟的同时节约索引的空间占用。
- 负责阿拉丁搜索以及用户、音乐等垂类搜索功能的从 Elastic Search 到 C++/Xapian/Hadoop 的架构迁移。初步探索搜索平台对大量异构垂类索引的架构支持。

Google 2018/03 - 2019/06

Google Cloud 微服务治理产品，HTTP 层全球负载均衡 *Sunnyvale, CA*

- 参与 GCP Traffic Director (托管的 Istio Pilot) 产品开发，作为控制层 API 接口对数据层的 Envoy proxy 提供 service mesh 相关的配置信息，以提供包括服务发现、负载均衡、安全验证等功能。设计并开发了 Envoy 配置信息缓存、地点信息上报、Ingress 安全配置等核心功能组件。
- 设计 Envoy proxy 在不同环境中的部署启动方案，环境包括 VM、GKE、原生 Kubernetes，以及它们分别在 GCP 和本地机房中的部署。
- 设计开发基于 Python 的集成测试框架，根据配置信息自动调用 Cloud API 生成相应测试环境，极大的简化了测试流程，提升了迭代效率。

Datera, Inc 2016/09 - 2018/02

分布式块存储数据层开发 *Sunnyvale, CA*

- 参与开发基于 Linux B-cache 和 iSCSI 的分布式块存储系统，对外提供对存储卷的多副本读写。维护数据层核心功能，包括数据副本检查、数据修复、不同存储节点的负载均衡等。
- 设计数据在 IO 操作中的加密解密、不同卷中相同数据块的存储去重、本地机房到 AWS 的远程备份等功能。

技术背景

编程语言	C/C++, Java, Python
大数据 & 分布式系统	Hadoop, Apache Flink