

杨清

2679frank@gmail.com | github.com/frank2679.io

个人总结

- 有五年以上异构平台 (CPU/GPU/DSP/NPU) 算子加速库经验, 包括 DNN, BLAS, FFT 等。深度参与 0-1 芯片软硬件研发;
- 有两年带团队经验, 团队规模 10 人左右;
- 有较强的软件工程能力 (项目落地), 很强的学习力, 坚持精益求精, 做精品。

教育经历/硕士研究生

台湾大学 | 电信研究所 | 学术型硕士研究生 2014.09—2017.01
WIFI 802.11ax 协议设计, 分别在 IEEE access 期刊, 及通信领域顶会 globalcom 上发表两篇论文。

- [1] **Yang, Hang**, D.-J. Deng, and K.-C. Chen, "On energy saving in ieee 802.11 ax," *IEEE Access*, vol. 6, pp. 47 546–47 556, 2018.
- [2] **Yang, Hang**, D.-J. Deng, and K.-C. Chen, "Performance analysis of ieee 802.11 ax ul ofdma-based random access mechanism," in *GLOBECOM 2017-2017 IEEE Global Communications Conference*, IEEE, 2017, pp. 1–6.

重庆大学 | 通信学院 | 工学学士 2010.09—2014.06
专业前 5%, 优秀毕业生, 获奖学金多次

技术能力

- 技术栈: 日常使用 C++, Python, cuda, 熟悉软硬件协同优化, zebu;
- 工作流: cmake, Linux, bash, Vim, Git, GitHub, JIRA, docker, markdown, jenkins, 英文工作环境。

工作经历/6 年

国内头部 GPU 厂商/2 年半 | 软件工程师/team leader/算子加速库 2021.1 至今
• 基于 GPGPU 平台从零开发 AI/HPC 领域加速库
• 带领 10 人左右团队

安防领域龙头企业/3 年半 | 软件工程师/AI 加速库/商业应用 2017.7—2020.12
• 异构平台高性能卷积神经网络 CNN 库开发, 调研评估芯片性能;
• 端到端负责算法侧项目, 深入梳理业务需求, 构建高效的应用方案, 加速智能算法落地;

项目经历

基于 C++ 开发 BLAS 加速库 | 0-1 开发维护 2021.11-至今
• lead 项目立项, 设计, 系统构建, 开发, 测试, 部署 CICD, 文档, 通过 JIRA 进行任务管理;
• 实现了算子注册框架, 日志系统, 核心算子开发;
• 核心算子 (GEMM/GEMV/reduce) 性能优化;
• 对软件产品的 SDLC 有了一定理解。

设计 DSL 开发 DNN 测试框架 | 基于 yaml 自定义算子用例表达 2022.3—2023.6
• 设计基于 yaml 的算子/图用例表达, 使得用例表达更加可读, 易于管理;
• 基于 gtest 设计 tester, 支持算子/图多种粒度的测试, 支持多个 backend 测试, 支持离线数据加载等;
• 对标业界主流精度对比方案, 支持多种 data type, op type, 特定硬件实现等弹性精度设置;
• 集成性能测试工具, 自动生成性能测试报告;
• 部署 CICD, 自动化守护 DNN 精度/性能;

GEMM 极致优化实现 | 基于 zebu 极致优化 2021.9—2021.10
• 手写汇编在 vcore 上实现极致性能 GEMM, 做到特定 shape 下利用率打满。
• 使用 zebu 查看 waveform, 优化到 cycle 级别。
• 基于 tcore 实现模拟 fp32 GEMM 算法, 也做到特定 shape 下利用率打满, 性能超过 A100。