

李明

☎ +86 17822001471 ✉ liming950814@gmail.com 🌐 <https://github.com/liming95> 🔗 liming95.github.io

个人简介

- 哈尔滨工业大学（深圳）博士研究生，研究方向为基于强化学习的后训练系统优化。
- 研究兴趣涵盖 GPU 加速计算、并行计算和软硬件协同设计与系统性能优化。
- 曾在中科驭数参与云原生 DPU 服务网格优化，实现端到端延迟降低 30% 以上。
- 参与 DPU 虚拟化与 FPGA 图计算加速研究，已取得 1 项专利、2 篇白皮书及多篇论文成果（如 FlatProxy、BitColor）。

教育经历

计算机科学博士（在读） 哈尔滨工业大学（深圳） 深圳 12/2025 - 至今

- 研究方向：大语言模型系统优化、基于强化学习的后训练（RL-based post-training）

计算机科学硕士 中国科学院计算技术研究所 北京 09/2020 - 06/2023

- 研究方向：计算机体系结构、网络系统
- 毕业论文：基于 DPU 的服务网格架构设计与优化
- 课程：高级计算机体系结构、数字电路设计、并行计算、机器学习

计算机科学与技术学士 天津大学 天津 09/2014 - 06/2015, 09/2017 - 06/2020

- 研究方向：软件工程与系统开发
- 课程：操作系统、编译原理、计算机网络、数据库系统
- 注：09/2015 - 09/2017 入伍

工作经历

研究实习 史蒂文斯理工学院 11/2024 - 12/2025

- GPU 加速的正则表达式匹配，具体为针对正则字符匹配在 GPU 上性能和可扩展问题的研究
- 调研 GPU 软硬件架构，矩阵乘优化和面向图计算（BFS）不规则负载的加速方法，具体可见相关的[总结报告 \[1,2\]](#)和[验证代码 \[3\]](#)
- 注：07/2024 - 03/2025 为 PHD 申请时间

软件工程师 中科驭数（YUSUR） 中国，北京 06/2023 - 06/2024

项目简介：

- 基于 DPU 构建服务网格，通过将 Envoy 代理卸载到 DPU，实现硬件加速的网络通信。

个人贡献：

- 设计并开发基于 Envoy 的代理框架支持公司自研 RPC 协议
- 通过简化数据路径，避免传输阻塞，降低端到端延迟 30% 以上
- 验证应用代理数据路径向用户态协议栈 VPP (Vector Packet Processing) 的卸载，包括 socket 层和传输层模式，其中 socket 层的卸载降低延时 50% 左右被部署于生产环境，传输层有 demo 验证和方案设计
- 部分参与通过 Docker 容器和 Kubernetes 在 DPU 侧部署 Envoy 代理

研究生 中国科学院计算技术研究所 中国，北京 06/2021 - 05/2023

- 研究基于 DPU 的服务网格体系结构：详细内容可见硕士毕业论文 [4]，同时参与 DPU 架构和测试相关白皮书撰写 [5,6]，发表论文 Flatproxy[7]
- 参与设计基于 SR-IOV 的 DPU 硬件辅助虚拟化方案：主要参与前期硬件架构的设计以及芯片后端 DFT 验证工作，发表一篇专利 [8]
- 其他：参与由 AMD-Xilinx 举办的 FPGA 定制化设计大赛，负责算法设计和部分硬件优化，获得二等奖以及发表论文 [9]

专利与发表

1. BFS 调研总结: <https://share.note.youdao.com/s/cDLF8mBN>
2. GPU 调研总结: <https://share.note.youdao.com/s/3swHfXtU>
3. 验证代码: https://github.com/liming95/cuda_algorithm/tree/main
4. 基于 DPU 的服务网格架构设计与优化. https://liming95.github.io/publication/master_thesis.pdf
5. 专用数据处理器 (DPU) 技术白皮书. https://liming95.github.io/publication/dpu_technology.pdf
6. 专用数据处理器 (DPU) 性能基准评测方法与实现. https://liming95.github.io/publication/dpu_benchmark.pdf
7. **Ming Li**, Wenyan Lu, Hanyue Lin, Jingya Wu, Yu Zhang, Guihai Yan. FlatProxy: A DPU-centric Service Mesh Architecture for Hyperscale Cloud-native Applications. arXiv, 2024
8. 一种资源调度的方法、装置及设备. <https://liming95.github.io/publication/patent-2022106929552.pdf>
9. Haishuang Fan, **Ming Li**, Jingya Wu, Wenyan Lu, Xiaowei Li, Guihai Yan. *BitColor: Accelerating Large-Scale Graph Coloring on FPGA with Parallel Bit-Wise Engines*. Proceedings of the 52nd International Conference on Parallel Processing (ICPP), 2023

获奖

- 中国科学院计算技术研究所优秀奖学金, 2022
- 第二届定制化计算挑战赛二等奖, 由中国计算机学会 (CCF)、AMD-Xilinx 与北京大学联合主办, 2021

其他

语言: 英文 (雅思 6.0)