

蔡经纬

中共党员

交叉信息研究院，清华大学

caijw21@mails.tsinghua.edu.cn

13258199537



教育背景

清华大学，交叉信息研究院，计算机系统结构，博士学位论文导师：马恺声

2021.9 - 今

- 研究兴趣包括基于 Chiplet 的 DNN 加速器架构和编译器设计，晶圆级硅光互联 DNN 计算系统设计，推荐系统加速等
- 发表论文五篇，其中以第一顺位作者发表 CCF-A 类计算机体系结构顶会论文两篇，DAC WIP Poster 一篇（共一）；获 HPCA2024 Distinguished Artifact Award (1/410)；ISCA2023、HPCA2024 Student Travel Grant
- 获长三角国际研发社区英才奖学金，院设奖学金，院社工优秀奖学金，研究生社会实践优秀奖学金
- 任交叉研二一班班长，交叉信息研会活动部部长（获评清华大学研会 2021-2022 小研之星称号），一二九主题教育活动交叉院领队（带队打破院史获综合二等奖，个人被评为“一二九”之星，优秀领队）

电子科技大学，电子学院，集成电路设计与集成系统，本科学位论文导师：黄乐天

2017.9 - 2021.6

- 国家奖学金 * 3 (当选 2021 年人民日报国家奖学金代表，全国共一百位)、成电杰出学生 (每年毕业生中遴选十位，类似清华特奖)、唐立新奖学金、感恩近现代科学家奖学金 (每年全校本研全体学生遴选十位，排名第一)、四川省优秀毕业生、电子科技大学优秀毕业论文奖、其余奖学金和资助性项目奖励共计十余项
- 加权均分 95.3/100 (打破学校纪录); GPA 3.99/4.0; 排名 1/643
- 任职班长四年 (带领班级连续三年获评校优秀班集体称号，个人获评校十佳班长、优秀团干等称号)；任职三下乡实践支队副队长 (获评成都市优秀三下乡团队，校实践特等奖，个人获评实践优秀个人)

实习经历

北极雄芯信息科技有限公司，编译团队主要技术负责人

2022.6 - 今

- 工作内容：带领编译团队将顶会论文的学术成果落地，搭建端到端的面向 12nm 大算力 DNN 加速器的可扩展的编译软件栈 (包含前端 IR 解析、调度优化、IR 生成优化、指令生成和优化等全流程)。该软件栈帮助该加速器大大提高了利用率和能效。

论文发表 (* 为共同一作)

- **Jingwei Cai***, Yuchen Wei*, Zutong Wu, Sen Peng, Kaisheng Ma “Inter-layer Scheduling Space Definition and Exploration for Tiled Accelerators” **ISCA 2023(CCF-A, 计算机体系结构顶会)**；获 **ACM Available、Functional、Reproducible Badge**；同时被 **Chinasys2023** 高分录用 (445)；该工作获得 **ISCA Travel Grant** 奖学金。该工作为第一篇对 Tiled/多核 DNN 加速器层间调度优化空间进行定义、探索和理解的文章，打破了传统的 layer-pipeline 和 layer-sequential 两种简单并行模式的垄断和巨大局限。该工作在组内创业公司得到完整落地，基于其搭建了端到端的编译框架 (包含前端 IR 解析、调度优化、IR 生成优化、指令生成和优化等全流程)。同时该工作受到来自学术界 (MIT、清华、北大、首尔国立大学、自动化所、计算所、华中科大等) 和工业界 (英特尔等) 的广泛关注和不同程度的使用，我们也为此提供了很多内部支持以提升该工具的影响力。
- **Jingwei Cai**, Zutong Wu, Sen Peng, Yuchen Wei, Zhanhong Tan, Guiming Shi, Mingyu Gao, Kaisheng Ma “Mapping and Architecture Co-exploration for Large-scale DNN Chiplet Accelerators” **HPCA 2024(CCF-A, 计算机体系结构顶会)**；获 **Distinguished Artifact Award(1/410)**，也是国内团队首次在 **HPCA** 获该奖项；获 **ACM Available、Functional、Reproducible Badge**；该工作获得 **HPCA Travel Grant** 奖学金。该工作为第一篇对 DNN Chiplet 加速器的架构和映射进行联合探索的工作，我们首次将成本考虑因素引入到 DNN Chiplet 加速器的架构探索中，开辟了新的研究方向。同时，本工作也是首次明确定义和深入探索 layer pipeline 空间映射的优化空间，为该领域的研究提供了清晰的方向和基础。通过对调度框架的精细探索，我们获得了许多颠覆传统认知的新见解，这些见解不仅拓宽了我们对 DNN Chiplet 加速器设计和优化的理解，也为未来的 Chiplet 加速器的研究和应用奠定了坚实的基础。
- Yuchen Wei*, **Jingwei Cai***, Mingyu Gao, Sen Peng, Zutong Wu, Guiming Shi, Kaisheng Ma “Discovering and Exploiting Untapped Buffer Resources in Many-Core DNN Accelerators” **DAC2024 WIP Poster** (CCF-A EDA 顶会)。本篇文章中，我们第一次观察到 Layer pipeline mapping 中存在本征的 Buffer Underutilized 现象，针对这一现象我们提出了新的 buffer 调度策略使能更多的 on-chip reuse 机会，取得了较大的性能和能效提升。
- Guiming Shi, Zhanhong Tan, Dapeng Cao, **Jingwei Cai**, Wuke Zhang, Yifu Wu, Kaisheng Ma “A 28nm 68MOPS 0.18J/Op Paillier Homomorphic Encryption Processor with Bit-Serial Sparse Ciphertext Computing” **ISSCC 2023** (芯片设计顶会)
- Guiming Shi, Yi Li, Xueqiang Wang, Zhanhong Tan, Dapeng Cao, **Jingwei Cai**, Yuchen Wei, Zehua Li, Yifu Wu, Wuke Zhang, Wei Xu, and Kaisheng Ma “PHEP: Paillier Homomorphic Encryption Processors for Privacy Preserving Applications in Cloud Computing” **HOT CHIPS 2023** (芯片设计顶会)

在研项目

- **DNN 加速器片上存储管理:** 对 DNN 加速器片上存储空间分配、地址管理进行优化，以最大程度高效利用片上存储，降低昂贵的 DRAM 访存，提高 DNN 加速器性能和能效。有一篇工作已经进入尾声，目标投稿 MICRO2024 (CCF-A 计算机体系结构顶会)。
- **基于硅光互连的晶圆级计算系统设计与研究:** 基于硅光互连的晶圆级 P 级计算系统，该项目基于科技部重点研发计划和浙大、西安光机所、重庆联合微电子等机构合作。我在其中作为组内代表，负责带领西安核心技术研究院工程师进行计算架构、编译器、芯片和封装的设计和开发。目前该项目相关 Chiplet 已经基本完成设计，准备流片。目标投稿 ISSCC 2025。