蔡经纬

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 - 2026.6

- **研究兴趣**：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
- **项目基金**：**中国科协青年人才托举工程博士生专项（清华叉院所有博士生仅一位），4 万元**
- **论文发表**：发表论文 6 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 3 篇，EDA 顶会 DAC WIP Poster 一篇（共一）；获**HPCA2024 Distinguished Artifact Award（1/410，中国首次）**；ISCA2023、HPCA2024 Travel Grant；
- **奖学金**：获**博士生国家奖学金**(清华叉院**所有**年级 CS 方向博士生中**每年仅一位**)，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
- **学术兼职**：JCST (CCF-B) 期刊审稿人，CCF-A HPCA2025, CCF-A 期刊 TCAD Subreviewer
- **主题 Talk**：DataFunSummit2024：AI 基础软件架构峰会；
- 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（带队打破叉院院史获综合二等奖，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)**; **排名 1/643**; GPA 3.99/4.0
- 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，编译团队技术负责人 + 芯片架构师 2020.6 - 今

- **芯片架构师**：
  - **职责**：在启明系列芯片的三代迭代过程中，我主要负责 NPU 多核互联架构、单计算核心微架构以及指令集架构的定义与设计。涉及的芯片包括启明 930（采用 RDL 封装，1 个 HUB Die 和 4 个 NPU Die，基于 TSMC 12nm 工艺，总算力达到 40TOPS）、启明 935（由 1 个 HUB Die 和 2 个 Side Die（NPU/多媒体）组成），以及启明 940（启明 935 的 AI 算力扩展和架构微升级版本）。
  - **解决的挑战**：由于我们开发了前沿的调度优化策略（ISCA2023, HPCA2024, DAC2024, HPCA2025 under review 系列工作），需要相应的芯片架构和指令级体系的支持，以充分发挥其潜力。因此，我主要负责将调度策略所需的硬件支持进行定义，并将其转化为架构规格和微架构设计。
  - **成果**：通过设计，芯片能够高效支持 chiplet 的可扩展特性，最大化发挥模块化架构的成本优势，实现了在主流自动驾驶神经网络中的高计算利用率（50%-90%，batch size 为 1）。
- **编译团队技术负责人**：
  - **职责**：作为编译团队的技术负责人，我带领团队将顶会论文中的学术成果成功应用于实践，**主导**构建了两代面向 12nm 大算力 DNN 加速器的可扩展编译软件栈。在开发过程中，我主要负责结合硬件特点设计各个优化模块的策略，并将这些策略分配给团队工程师进行实现。同时，我负责解决技术难题和代码审核，并主导软硬件耦合，将软件侧发现的问题或需要支持的功能同步至芯片微架构优化。
  - **软件栈特点**：第二代软件栈在吸取第一代经验的基础上进行了底层重构，涵盖了从前端 IR 解析、算子融合与图优化、数据预取优化、片上存储地址优化、IR 生成优化、指令生成与优化到指令级仿真器的完整流程。软件栈采用模块化设计，便于各优化模块的升级与调优，并支持通过 chiplet 组合的不同算力版本的编译需求。
  - **成果**：该编译软件栈成功支持了 ONNX 中超过 90% 的算子，并在主流自动驾驶神经网络中实现了 50%-90% 的计算利用率（batch size 为 1）。
  - **当前工作**：目前，团队正全力支持大语言模型（LLM）网络的编译和调优工作，进一步提升性能与效率。

论文发表 (* 为共同一作)

- **Jingwei Cai**, Zuotong 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 IEEE***

Available、***Functional***、***Reproducible Badge***；**该工作获得** ***HPCA Travel Grant*** **奖学金。**

- **研究创新**：本工作提出了首个针对大算力 DNN chiplet 加速器的映射和架构联合探索框架，并首次将成本考虑因素引入到 DNN Chiplet 加速器架构探索中，填补了该领域的研究空白。本工作还首次明确定义并深入探索了 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，本工作在 Chiplet 粒度等架构问题上和空间映射调度层次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 加速器研究和应用奠定了坚实的基础。
- **亮点**：本研究获得了 CCF-A HPCA 2024 的“Distinguished Artifact Award”（1/410），这是中国团队首次在 HPCA 中获得此殊荣。

- **Jingwei Cai***, Yuchen Wei*, Zuotong 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。**
  - **研究内容**：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。
  - **产品化落地**：该工作在组内创业公司（北极雄芯）得到了**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。
  - **学术影响**：该工作受到学术界（MIT、清华、北大、佐治亚理工、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，并在不同程度上得到了使用。我们为此提供了大量内部支持，进一步提升了该工具的影响力。
  - **亮点**：佐治亚理工的体系结构专家 Tushar Krishna（Gem5 和 Garnet 作者，Google Scholar 引用次数 16000+，ISCA 2023 PC Chair）将我们的框架纳入其教授课程的一部分，供学生进行学习和使用。

- **Jingwei Cai**, Xuan Wang, Mingyu Gao, Sen Peng, Zijian Zhu, Yuchen Wei, Zuotong Wu, Kaisheng Ma “Identifying, Exploring, and Understanding the DRAM Communication Scheduling Space for DNN Accelerators”***HPCA 2025 (CCF-A, 计算机体系结构顶会)***;
  - **研究创新**：随着 DRAM 带宽和计算密度的鸿沟越来越大，优化 DRAM 通讯已经成为编译调度优化的核心任务之一，本工作针对 DRAM 通讯进行了优化，主要包括细粒度层融合优化以及数据预取以及延迟发送优化。细粒度层融合优化相较于传统的算子融合可以更加节约片上缓存占用，数据预取和延迟优化则是之前被很大程度忽略的优化维度，本文对其进行了全方位的分析和优化。最终本文将两种优化维度进行了统一表示，开发了一个框架进行联合探索优化，和 SOTA 的 ASPLOS COCCO 相比我们可以提高 2.44 倍的性能并降低 48.2% 的能耗。
  - **亮点**：该文章对应框架已完成对一款大算力自动驾驶芯片的针对性编译器开发，完全落地。现已回片，预计明年**量产**。

- Yuchen Wei*, **Jingwei Cai***, Mingyu Gao, Sen Peng, Zuotong 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 pipeine 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 (均为负责人或主导者)

- **LLM 推理服务加速研究 2024.5-今**：MOE 分布式推理加速，研究适配 MOE 模型的多维并行策略以及计算传输掩盖策略。

- **基于 3D DRAM 技术的 LLM 推理加速器架构和编译研究 2024.5-今**：LLM 推理将带宽的需求推到前所未有的程度，而近些年随着 3D 工艺的进步 3D DRAM 技术逐渐成熟，但是相较于传统 DRAM，3D DRAM 先天的分布式特征对架构设计和调度都提出了新的要求，如何设计合理的粒度以及调度方式充分发挥 3D DRAM 的带宽潜力是一个十分重要的问题，也是本项目的中心。

- **DNN 加速器多网络调度优化 2024.01-今**：针对云服务、自动驾驶等 multi-DNN 和 multi-tenant 的场景，创新性的以 layer-block 作为粒度进行调度优化，相较于传统的以层为粒度或者以网络为力度可以在满足动态指标要求的前提下尽可能挖掘静态优化的潜力，目前该工作正在进行中，目标投稿 ISCA2025 （CCF-A 计算机体系结构顶会）；**同时该工作将在北极雄芯公司产品化落地**。

- **基于硅光互连的晶圆级计算系统设计与研究 2022.10-今**: 自 2022 年起，我作为组内代表，深度参与了“十四五国家重点研发计划重点专项”——晶圆级硅光交换互连片上计算系统项目。该项目利用先进封装技术，集成了 48 个计算芯粒和 4 个硅光交换芯粒，旨在实现 P 级计算能力的晶圆级计算系统，技术挑战极大。
  - **职责**：我领导由清华大学和西安核心院的研究生及工程师组成的团队，全面负责计算系统架构和编译器的研发，并在整体封装架构设计中发挥了关键作用。
  - **创新技术方案**：在项目中，我提出了一系列创新性技术方案，成功推动了项目的顺利进展，并有效应对了多个核心挑战。特别是基于硅光交换互联的特性，我提出了异构芯粒计算架构和多层次混合并行调度策略，充分发挥了计算芯片与硅光互联芯片的各自优势，显著提升了系统的整体性能。
  - **项目进展**：目前，该项目已进入最后的流片阶段，相关成果预计将在明年投稿至 *Nature Photonics* 以及芯片领域的顶级会议 *ISSCC*。