

吴坎

☎ (+86) 18756280097 · ✉ wukan3@mail2.sysu.edu.cn · 🌐 <https://wu-kan.cn> · 🌐 wu-kan

🎓 教育

中山大学, 广州

硕士 (在读), 计算机科学与技术

2021 年 9 月 ~ 2024 年 6 月 (预计)

- 导师: [张献伟@国家超级计算广州中心](mailto:zhangxianwei@nsc-gz.com)
- 研究方向: sw-hw co-design; GPU arch

学士, 计算机科学与技术

2017 年 9 月 ~ 2021 年 6 月

- 专业必修: 计算机组成原理、操作系统原理、计算机网络、编译器构造实验 (1/79)
- 超算班限修: 超算原理 (1/39)、并行与分布式计算、高性能计算、多核程序设计 (1/84)

👤 经历

中山大学 计算机学院

2021 年 9~12 月, 2022 年 2~6 月 (预计)

助教 编译原理 计算机组成原理

- 独立设计编译原理实验 [SYsU-lang](#), 已应用于课堂教学
 - 使用不超过 3000 行代码, 基于 Clang/LLVM 实现精简 C 语言编译器
 - 教学效果良好, 实验发布一周, 已有同学达到 `clang -O3` 的 25%, 加速比超过 100%
 - [《SYsU-lang: 步步为营, 构建编译实践大局观》](#) 正在投稿 2022 年中国高校计算机教育大会
- 发起并参与设计组成原理实验 [YatCPU](#) (“逸芯”)
 - 一个使用 Chisel 编写的五周期流水线 CPU, 当前支持指令集为 RV32I

字节跳动

2021 年 2~3 月, 2021 年 7~8 月

算法引擎实习生 AI Lab Speech & Audio Team (Beijing)

- Panther 推理引擎开发的 GPU 后端
- 开发定制 TransformerDecoder 算子, 同时编写对应单元测试、图匹配脚本
 - 相关模型在 NVIDIA T4 上推理加速 1.6× (80ms→50ms)
- 基于 CUTLASS 的量化 MHA 算子 (batchGemm-Softmax-batchGemm)
 - 序列长度较短时, 相较基于 cublasLt 的原始实现加速 7 倍以上

国家超级计算广州中心

2020 年 3 月 ~ 2021 年 6 月

队长 中山大学超算队

- 在 ASC 决赛中
 - 暴力调整主板接线实现激进功耗控制, 在三千瓦功耗限制下打破 Linpack 纪录并提升 57%
 - 负责量子线路模拟器赛题 QuEST 在多 GPU 平台上的性能优化
- 在队内定期开展技术分享
- 作为首个 (一年内) 在国内所有超算竞赛获前三名的选手, 受邀参与 IPCC 赛前讲座
 - [超算竞赛备赛指南](#), [超算基本原理与方法论](#), [现代处理器优化概论](#)
 - 取得较好反响, 单次中直播收到反馈问卷上百份

📁 项目

RollBin: Reducing code-size via loop rerolling at binary level

2022 年 2 月 ~ 3 月

Tianao Ge, Zewei Mo, **Kan Wu**, Xianwei Zhang, Yutong Lu [LCTES'22](#)

- 本文提出一种二进制代码优化手段 RollBin, 在实验中相比已经存在的提升 31% ~38%
- 参与论文思路讨论、实验验证、数据整理与可视化、论文撰写

面向新一代神威超级计算机的 Vision Transformer 模型优化

2021 年 8 月 ~ 9 月

SACA (*Sunway Accelerate Computing Architecture*) “神威杯”参赛作品

- 在新一代神威超算上优化 SWPyTorch 使用 256 节点训练 ViT 模型的过程
- 前向过程加速上千倍，反向与 `optimizer.step()` 过程加速上百倍
- 算例规模下六核组 SGEMM 峰值性能超过 8.2 TFlops，达到理论性能的 59.24%

面向国产异构处理器的 HPL-AI 基准实现及优化

2020 年 11 月 ~ 2021 年 2 月

AscendCL/CUDA 本科毕业设计

- 首个开源的多节点 HPL-AI 基准测试，基于 hpl-2.3 实现，相较前者达到 2.59x 加速比
- 使用 NPU (昇腾 910) 的半精度算力加速 Linpack 求解，然后通过数值迭代提升精度
- 在 X86、ARM、GPU、NPU 等多种环境上通过正确性、可扩展性验证

♥ 荣誉

二等奖 (3/130+), ACM 中国-国际并行挑战赛 (ACM-IPCC)	2021 年 11 月
二等奖 (3/100+), “神威杯”并行应用挑战赛 (CPC)	2021 年 10 月
CCF 优秀大学生 (全国共 94 名, 中山大学共 1 名), 中国计算机学会 (CCF-ECA)	2021 年 9 月
优秀本科毕业论文 (Top 5%), 中山大学	2021 年 5 月
最高性能奖 & 一等奖 (3/400+), 领队, 世界大学生超级计算机竞赛 (ASC)	2021 年 5 月
金奖 (4/98), CCF 大学生计算机系统与程序设计竞赛 (CCSP), 华南赛区	2020 年 10 月
一等奖学金 & 学科竞赛一等奖学金 (Top 5%), 中山大学	2020 年 10 月
银牌 (2/100+), “英特尔杯”并行应用挑战赛 (PAC)	2020 年 9 月
一等奖 (3/unknown), 中科院“先导杯”并行应用挑战赛 (CAS-PRA)	2020 年 8 月
Top 0.05% (6/11395), CCF 计算机软件能力认证 (CCF-CSP)	2019 年 12 月
银牌, 国际大学生程序设计竞赛 (ICPC), 徐州站	2019 年 11 月
金奖 ×2, 中国大学生程序设计竞赛 (CCPC), 广东赛区	2018 年 5 月, 2019 年 5 月

⚙️ 技能

- CUDA/HIP/ROCm/MPI/OpenMP/Pthread
 - 使用 CUDA 实现 SGEMM, 达到 NVIDIA V100 理论算力 92.1%
 - 同情况下 CUTLASS 库为 92.6%, cuBLAS 库为 97.0%
- Linux
 - 多台超算 (含天河、神威、曙光) 使用、开发经验, 超算队集群、学院多台集群运维
 - 在云服务器上通过 docker 搭建一系列自用服务, 如 [code-server](#)、[gitlab](#)