

# 张浩源

(+86) 186-1076-7285 · zhanghaoyuan@cnic.cn · <https://microzhy.github.io>

## 教育背景

中国科学院大学, 博士研究生, 高性能计算方向, (预计 2026 年 6 月毕业) 2019.9 - 2025  
 河海大学, 学士学位, 工程力学专业, 排名 1/70 2015.9 - 2019.6

## 技术能力

- 了解异构芯片 (NVIDIA/AMD GPU, 国产众核) 架构特点及软硬件优化策略, 具备实际开发调优经验。
- 熟悉 CUDA 编程和性能优化技术, 包括 **TensorCore**、Memory Coalescing、Bank Conflict、Double Buffer。
- 掌握 Nsight Compute 等 profile 工具, 能够独立定位, 分析和解决性能问题。
- 掌握 C/C++、Python 语言, 具备扎实的算法基础和良好的编程风格。
- 掌握 MPI/OpenMP 编程模型和优化策略。
- 了解大模型推理 Pipeline 和常用优化手段; 了解主流推理框架 vLLM 并具备推理性能优化经验。

## 项目经历

国家重点研发计划 | 跨域异构环境 CAE 软件高效求解方法 2021.2 - 2024.9

- 项目背景:** 聚焦于优化 NVIDIA GPU 和国产类 GPU 加速的线性解法器, 深度适配硬件架构。
- 技术创新:** 提出基于 **Tensor Core** 的预处理算法和稀疏矩阵运算优化策略, 针对低算术强度问题设计访存高效的数据结构; 提出块稀疏矩阵-向量乘加 (BSpMVA) 高性能算子, 作为子区域预条件在 NVIDIA A100 GPU 上相比 cuSPARSE 实现平均 **4.63** 倍加速。
- 解决效果:** 整体性能相比最流行的科学计算工具库 PETSc 提升超过 **10** 倍。

美团校招实习 | 计算与智能平台部-模型压缩与加速组 2025.6 - 至今

- 项目背景:** 面向大模型 RL 训练采样场景, 解决显存占用瓶颈与长序列延迟问题。
- 技术创新:** 在 vLLM 框架中支持 **INT8 KVCACHE** 量化功能。
- 解决效果:** 在 NVIDIA H800 GPU 上, 在 MOE 48B 模型上实现 **20%** 性能收益。

## 学术论文

- [ICCD'25, CCF B] **Haoyuan Zhang**, Yaqian Gao, Xinxin Zhang, Jialin Li, Runfeng Jin, Yidong Chen, Feng Zhang, Wu Yuan, Wenpeng Ma, Shan Liang, Jian Zhang, Zhonghua Lu. *FlashMP: Fast Discrete Transform-Based Solver for Preconditioning Maxwell's Equations on GPUs*.
- [ICCD'24, CCF B] **Haoyuan Zhang**, Yidong Chen, Wenpeng Ma, Wu Yuan, Jian Zhang, Zhonghua Lu. *MIST: Efficient Mixed-Precision Preconditioning Through Iterative Sparse-Triangular Solver Design*.
- [CCF THPC'24, CCF C] **Haoyuan Zhang**, WenPeng Ma, Wu Yuan, Jian Zhang, Zhonghua Lu. *Mixed-precision block incomplete sparse approximate preconditioner on Tensor core*.
- [Frontiers of Data & Computing'24] **Haoyuan Zhang**, Wenpeng Ma, Wu Yuan, Jian Zhang, Zhonghua Lu. *Implementation of CCFD-KSSolver Component for GPU Architecture*.
- [SC'24, CCF A] Yidong Chen, Chen Zhang, Rongchao Dong, **Haoyuan Zhang**, Yonghua Zhang, Zhonghua Lu, Jidong Zhai. *MIXQ: Taming Dynamic Outliers in Mixed-Precision Quantization by Online Prediction*.
- [ICPP'24, CCF B] Runfeng Jin, Wenhao Liang, **Haoyuan Zhang**, Yinxuan Song, Zhen Luo, Haibo Ma, Yingjin Ma, Zhong Jin. *PASCI: A Scalable Framework for Heterogeneous Parallel Calculation of Dynamical Electron Correlation*.

## 竞赛获奖

- [**全国一等奖**] 第 11 届并行应用挑战赛 (PAC2024), 队长; 负责并行流场模拟软件在国产异构平台 (DCU) 上的多层次片上内存体系优化, 通过合并访存、细粒度预取等技术将程序热点加速 **27** 倍。
- [**全国三等奖**] 第七届国产 CPU 并行应用挑战赛 (CPC2023), 队长; 在国产超算众核架构上, 使用 Athread 编程模型移植稀疏迭代解法器, 通过算子融合、DMA 等技术在单核组上取得 **34** 倍加速比。
- [**全国三等奖**] ACM 中国国际并行计算挑战赛 (IPCC2022), 队长; 在 **AMD CPU** 平台上, 采用串行算法优化 (消除冗余计算, 访存局部性) 与并行算法 (最小生成树, BFS) 设计, 取得显著加速。
- [**全国三等奖**] 第三届先导杯计算应用大奖赛 (PRA2022), 队长; 针对特征值求解问题, 在 **AMD GPU** 架构上通过多线程并行、循环拆分等技术加速, 较原版 rocSOLVER 接口加速 **35.8** 倍。