

# 孙忠奥

求职岗位：推理框架优化/算子优化/高性能计算工程师

年 龄： 25岁

性 别： 男

政治面貌： 共青团员

电 话： 13921394067

邮 箱： jlucs\_sza@163.com

Github/Gitee： ASCII-S



## 教育背景

2023-09 ~ 2026-06

吉林大学(985)

计算机专业 (硕士)

主修课程：人工智能原理，最优化理论，分布式系统；CCF图计算挑战赛二等奖(队长)；合作产出JCST(CCF-B)一篇；

2018-09 ~ 2022-06

江苏科技大学

高分子材料与工程 (本科)

成绩排名5/38；主修课程：概率论(96),电工电子技术(98); 英语六级(575);人民奖学金;生活委员;

## 专业技能

- 精通C++,掌握Python,Shell;熟悉Linux环境;掌握CUDA,MPI,Openmp;
- 熟悉CPU/GPU架构;掌握常用性能优化方法;
- 熟悉深度学习基本原理,Transformer架构;熟悉模型量化技术(PTQ,QAT);掌握vllm及相关推理优化技术如KVCache,FlashAttention,Pageattention;熟悉分布式推理(TP,PP,DP);
- 熟悉Git/Markdown/CMake/Makefile；掌握常用调试工具，性能分析工具

## 实习经验

2025-06 ~ 2025-09

国家超算无锡中心

实习工程师

- 面向Sunway芯片和LoongArch最新芯片3C6000D,设计性能指标,基于perf,iostat,numastat等系统工具编写性能测试脚本,以海洋海浪模式(WW3)为例分析机器计算,内存,io扩展性.识别并对各架构的关键拐点和瓶颈进行归因分析,撰写报告为应用移植和性能优化提供方向
- 基于linux系统调用,编写面向代码行的简易性能分析工具,快速定位并行程序(NDGTD)瓶颈所在代码行.

2023-05 ~ 2023-09

国家超算无锡中心

性能优化实习工程师

- 调研性能优化方法.通过优化访存顺序,内存布局,simd,多线程等技术加速矩阵乘算法,最终在cpu平台上得到相当于openblas数学库107%的浮点计算性能
- 通过矩阵分块,合并访存,避免bank冲突等性能优化技术,优化gemm在gpu上的性能表现

## 项目经验

2024-09 ~ 至今

大规模分步式图计算框架(国家重点项目)

研究助理

- 参与移植基于MPI的分布式图遍历算子BFS,Kcores使得算子能够并行执行
- 调研图计算框架,设计分布式异构计算模块,赋能图算子的在多机器间异构计算(cpu,gpu,专用加速卡)能力,并且初步支持设备级别的容错处理.

2023-09 ~ 2025-01

HPC程序可修复性分析

硕士课题

- 设计程序插桩模式,静态分析程序汇编指令,筛选程序的软错误故障位点,实现对软错误的故障模拟;
- 设计程序在多种故障信号下的启发式修复方式,使得程序崩溃后60%机概率恢复运行
- 基于python的pexpect库与GDB工具,编写脚本自动化交互程序并收集程序状态

2025-09 ~ 2025-10

轻量级推理框架开发

独立开发

参考vllm架构及源码, 基于python实现了qwen3-0.6b模型的离线推理, 应用优化技术包含连续批处理, paged kvcache