CHEN Han

Last update: Augest 20, 2024

CONTACT INFORMATION

Tel: +86-13305014345
Tel: +65-89420214
Wechat: concyclics
✉ E-mail:chenhan@u.nus.edu
Github:www.github.com/Concyclics
Address: 05-02, West Coast Residental Village, 127371, Singapore

EDUCATION

National University of Singapore, Singapore. 2023–2025(Expected)
- Master of Computing, Computer Science Specialization. GPA: 4.3/5.0

South China University of Technology (SCUT), Guangzhou, China. 2019–2023
- B.Eng., Software Engineering. GPA: 3.6/4.0

PRIZES AND AWARDS

- **Excellent Degree Dissertation of South China University of Technology** 2023
- **Honorable Mention** in Mathematical Contest in Modeling 2023
- **National Scholarship** 2022
- ***Bronze Medal (46th)*** in ICPC Asia-East Continent Final(Xi An) 2022
- ***101/1608*** in CCF-DBCI Competition of "Small Sample Data Classification" 2022
- ***Silver Medal (46th)*** in ICPC Asia Regional Contest(Ji Nan) 2021
- ***44/3567*** in CCF-DBCI Competition of "Recognition of figure skaters' skeleton points based on Paddle" 2021
- ***First Prize*** in National Olympiad in Informatics in Province(NOIP) 2017

RESEARCH EXPERIENCE

- ***Research Assistant*** in **National University of Singapore**
  Mentor: Prof. Bingsheng HE May-Sept. 2024
  - Research on efficient inference for large language models(LLMs).
  - Deploy a Qwen-72B model inference with tensor parallel and 4-bit quantized KV cache and speculative decoding method on 4 NVIDIA A100 GPU system for partner corporation 4 Paradigm.
  - Design a new 2-bit KV Cache quantization for LLMs, which can reduce the memory usage by 4-8x with more than 90% accuracy retention on complex tasks like GSM8K and Code Completion.
  - On submission to ICLR 2025.
- ***Internship*** in **SG Digital Trust Lab, Singapore Research Center, 2012 Laboratory**
  Mentor: Dr. Tao HUANG Jan.-May 2024
  - Research on high-performance symmetric encryption algorithm and SIMD optimization with AES instructions.
  - Optimize LOL-MINI-NMH algorithm with scroll array and XOR fusion feature of KUNPENG 920 processor to improve the performance from 7.1Gbps to 8.5Gbps.
  - Realize a new stream cipher algorithm with 49Gbps performance on KUNPENG 920 processor with the same AES instructions involved(3:1) as SOTA method Rocca which run at 38Gbps on the same processor, and also achieve a 1.3-1.4x performance over Rocca on other ARM processors.
  - On submission to FSE 2025.

- ***Symmetric Matrix Solving Algorithm Parallel Optimization for ARM Architecture***
  Mentor: Prof. Deyou TANG May-Dec. 2022
  - Optimize and parallel Bounded Bunch-Kaufman Algorithm(*sysv_rk subroutine of LAPACK) for solving symmetric matrix on ARM server processor with NEON instruction set and openMP.
  - Implement a parallel column reordering method in row swap of solving symmetric matrix to enhance memory access locality for column major matrix for better cache hit rate and parallelism, achieving a performance improvement from 320Gflops to 580Gflops.
  - Implement the same optimization on Skylake Intel processor and achieve 2-5x multi-core speedup than MKL library for *sytrs_3 subroutine of LAPACK.
  - Awarded as the Excellent Degree Dissertation of South China University of Technology.

Technical Skills

- *English*: IELTS(6.5), CET-4, CET-6.
- *Programming Languages*: C/C++, Fortran, p4-16, Python, SQL, LaTeX.
- *Technical Skills*: openMP, SIMDs(NEON, AVX512), MPI, PyTorch, CUDA.
- *TestDemo Certificate*: C++, TOP 10%, LINUX, TOP 10%, PYTHON, TOP 10%.
- *Kaggle Certificate*: Data Visualization, Intro to Machine Learning, Intro to Deep Learning, Intro to Game AI and Reinforcement Learning.

Exchange Experience

- **Online Academic Program on Machine Learning, McGill University** Jan.-Feb. 2022

陈涵

最近更新：**2024年8月20日**

联系方式

Tel: +86-13305014345
Tel: +65-89420214
微信: concyclics
✉ E-mail:chenhan@u.nus.edu
Github:www.github.com/Concyclics
地址: 新加坡 West Coast Residental Village 05-02, 127371

教育经历

新加坡国立大学, 新加坡 2023–2025(预计)
- 计算机科学硕士, 计算机科学方向. GPA: 4.3/5.0

华南理工大学, 广东省广州市 2019–2023
- 工学学士, 软件工程专业. GPA: 3.6/4.0

获奖荣誉

- **华南理工大学本科优秀毕业设计(论文)** 2023
- **二等奖** 美国大学生数学建模竞赛 (MCM/ICM) 2023
- **铜牌** 第46届ICPC国际大学生程序设计竞赛亚洲区决赛 2022
- ***101/1608*** CCF-DBCI "小样本数据分类算法" 竞赛 2022
- **国家奖学金** 2022
- **银牌** 第46届ICPC国际大学生程序设计竞赛(济南站) 2021
- ***44/3567*** CCF-DBCI "基于飞浆实现花样滑冰选手骨骼点识别" 竞赛 2021
- **一等奖** 全国青少年信息学奥林匹克联赛(NOIP) 2017

项目经历

- **科研助理**: 新加坡国立大学 2024/05–2024/09
  导师: 何丙胜教授
  - 研究大型语言模型(LLMs)的高效推理。
  - 通过张量并行和4位量化KV缓存, 以及推理过程中的投机采样的方法, 在4 NVIDIA A100 GPU系统上为合作公司第四范式部署了Qwen-72B模型推理。
  - 设计了一种新的2-bit KV缓存量化方法, 可以在复杂任务如GSM8K和代码补全中减少4-8倍的内存使用, 并保持90%以上的准确性。
  - 该项目计划投稿至ICLR 2025。
- **实习生**: 华为2012实验室新加坡研究所数字信任实验室 2024/01–2024/05
  导师：黄涛博士
  - 研究利用SIMD指令集实现的高性能的流式对称密码算法。
  - 通过滚动数组优化和鲲鹏920处理器的异或指令融合特性, 将LOL-MINI-NMH算法的性能从7.1Gbps提升到8.5Gbps。
  - 实现了一种新的对称流密码算法, 与当前SOTA算法Rocca相比, 在使用相同比例AES指令(3:1)的情况下, 在鲲鹏920处理器上达到了52Gbps的性能, Rocca算法在该处理器上性能为38Gbps，同时在其他型号ARM处理器上也有约30-40%的性能提升。
  - 该项目计划投稿至FSE 2025。
- **对称矩阵函数求解*BBK*算法的并行优化** 2022/04-2022/12
  导师: 汤德佑教授
  - 在ARM处理器上利用NEON指令集和openMP对Bounded Bunch-Kaufman算法(LAPACK库*sysv_rk 函数)进行并行优化。
  - 实现了一种并行列重排方法, 在列优先矩阵的行交换中改进访存局部性, 使得缓存命中率和并行性能得到提高, 在鲲鹏920-6426处理器上的单精度性能从320Gflops提升到580Gflops。

- 将该方法移植到Intel Skylake处理器上, 对比MKL库的*sytrs_3函数, 实现了2-5倍的并行性能提升。
- 该项目获评华南理工大学本科优秀毕业设计。

专业技能

- 英语认证水平: CET-4, CET-6, IELTS(6.5).
- 编程语言: C/C++, Fortran, p4-16, Python, SQL, LaTeX.
- 编程技能: openMP, SIMDs(NEON, AVX512), MPI, PyTorch, CUDA.
- *TestDemo* 编程技能认证: C++, TOP 10%, LINUX, TOP 10%, PYTHON, TOP 10%
- *Kaggle* 课程认证: 数据可视化, 机器学习, 深度学习, 强化学习

交换经历

- **机器学习线上访学项目, 麦吉尔大学** 2022/01–2022/02