姜前辰

(+86) 135-0758-8162 · 福州大学 (211) · 计算机科学与技术 · 1569978990@qq.com · 23 岁

我是姜前辰, 毕业于福州大学计算机与大数据学院, 曾担任校超算团队负责人, 获得过 ASC 世界大学生超级计算机竞赛一等奖等荣誉。现就职于太初电子科技有限公司的推理算子组, 主要负责在异构众核国产芯片上开发大模型推理算子, 参与完成了适配 vLLM 推理框架的相关算子开发以及主流新模型的算子适配工作。希望能有更多机会学习新知识并实践想法。

教育背景

2020.09 – 2024.06 | 福州大学(211) · 计算机与大数据学院 · 计算机科学与技术 · GPA: 3.3 / 4 · Rank: 30% · CET6

工作经验

北京科学智能研究院 **高性能优化实习生** 2024 年 03 月 – 2024 年 06 月

- **软件内存跟踪优化**: 利用 Vtune 工具, 动态库打桩等方式绘制软件 CPU 版本的内存峰值曲线，采用分块计算的思想优化内存消耗高的部分。
- **CUDA 算子融合**: 利用 Nsght-system 和 Nsight-computer 工具分析 CG 对角化计算程序热点, 将单 kernel 小算子及访存瓶颈模块采用算子融合的方式优化。

太初电子科技有限公司 **推理算子工程开发** 2024 年 07 月 – 至今

- 负责**大模型推理融合算子**的开发，参与适配异构众核加速卡的 **vLLM 推理框架相关算子开发**（参与开发如 Paged Attention, Continuous Batching, FFN, Top_k/p 等相关核心算子）。
- 参与算子的**高性能优化**：设计相关模型权重 layout 实现高效访存策略，设计并行流水策略，编写高性能 SIMD 代码，并进行汇编级别的代码优化。
- 紧跟 LLM 模型的技术进展，完成如 DeepSeek 模型 MLA 算子的开发与接入, 如 GPT-OSS 模型相关的 MXFP4 量化权重计算, sliding-window&sinks 等相关模块开发。

项目经历

支持 Baidu Comate 代码大模型相关异构众核算子开发 2024 年 10 月 – 2024 年 12 月

1. 参与完成 Paddle 算子 block_multihead_attention 非首字模块在异构众核加速卡上的适配
   - 设计 block 内的权重私有排布格式，实现 KV_Cache 的 DMA 访存带宽**达理论峰值的 75%+**。
   - 对非首字 Attention 计算 中的 qk 模块性能优化，利用 **ACE** (矩阵乘法，脉动阵列) 代替 SIMD 计算组件，在模型 q_per_kv > 4 的情形下实现算子性能超 20% 的加速提升。
   - 算子功能泛化：开发适配异构众核芯片的 **Flash Decoding 计算** (非首字计算过程中将 k, v 分块，每个块分别与 q 做 FlashAttention, 分块之间并行通过 reduce 得到最终结果)，实现对**无限长输入序列**的支持。并利用组内自研的预设**二幂分配方法**进行优化, 实现计算资源的高效利用。
2. 开发 Paddle 采样算子 (top_p_sampling_reject)
   - 设计多核架构下的**大小堆 Topk 选择**方案，相比于传统排序算法实现了 4.2 倍的性能提升。

大模型推理算子 Weight Only 量化支持 2025 年 03 月 – 2025 年 06 月

在异构众核架构的国产芯片上完成对 INT8, INT4, MXFP4 计算精度的支持

1. 由于芯片硬件不支持量化精度数据格式的计算，参与设计开发**离线量化，在线反量化的 GEMM 计算模块**（计算流程：原始 fp16 权重 => 转化成高性能重排的量化私有格式权重 => 在线反量化为 fp 16 权重 => fp16 GEMM 计算）
2. 参与设计开发权重转化接口 (WeightPermute), 实现权重 DMA 访存带宽**达理论峰值的 75%+**。
3. 参与 GEMM 模块异步器件的**多级流水设计**，在非首字属于 memory-bound 的情况下，使用 DMA 传输耗时掩盖其他异步计算组件的耗时。
4. 参与设计开发**高性能 SIMD 反量化计算**，利用查表指令以及相关数值转化的特性压缩 SIMD 计算指令，充分用指令流水以及指令双发射的特性重排指令，实现**仅需 245 个 cycle** 完成对 32*32 (1024) 个元素的量化权重单元反量化计算。
5. 开发算子接入主流的模型官方量化权重（包括 **GPTQ, AWQ 量化**，以及 **MXFP4 量化精度**），其中 4 bit 量化相比于 fp16 精度,实现性能达 1.6 倍左右的加速提升，显存消耗节省超 3.4 倍; 8 bit 量化相比于 fp16 精度， 实现性能达 1.7 倍左右的加速提升，非首字等效带宽达理论峰值的 68%+。

获奖情况

ASC 世界大学生超级计算机竞赛一等奖 2023 年 05 月

太初元碁 天道酬勤奖 公司 AI 社区校招生算子开发第一名 2024 年 08 月

太初元碁 新锐突破奖 公司年度优秀新人 2025 年 03 月