PR #7007 完整报告

PaddlePaddle/FastDeploy

[Optimization] optimize fused_swiglu_fp8_quant_kernel

合并时间：2026-03-27 16:10

原文链接：http://prhub.com.cn/PaddlePaddle/FastDeploy/pull/7007

执行摘要

本 PR 优化了 GPU kernel fused_swiglu_fp8_quant_kernel，通过共享内存加速 token 映射、缓存专家范围和增加向量化尺寸，在特定测试配置下实现 20%-30% 的性能提升，适用于混合专家模型的 SwiGLU 层 FP8 量化计算。

功能与动机

优化动机源于 kernel 性能与 sum(token_nums) 相关，现有实现存在计算瓶颈。PR body 展示在 B 卡上测试，GROUP_NUM=20、GROUP_SIZE=4096 配置下，优化前性能较低，优化后显著提速，旨在提升推理效率。

实现拆解

仅修改文件 custom_ops/gpu_ops/fused_mask_swiglu_fp8_quant_kernel.cu，关键改动：

- 共享内存前缀和：使用共享内存构建 token_nums_per_expert 的前缀和表，加速 token 到专家的映射查找。
- 向量化提升：将向量大小从 4 增加到 8（VEC_SIZE = 8），提升内存带宽利用。
- 专家缓存：引入 cached_expert、cached_cumsum_lo、cached_cumsum_hi 变量，缓存当前专家范围，避免重复二进制搜索。
- 预计算常量：在循环外预计算 inv_fp8_max 等常量，减少重复计算开销。

评论区精华

review 过程无实质性讨论，reviewer freeliuzc 直接批准，表明变更被认为可靠或简单。

风险与影响

风险：

- 共享内存大小依赖 group_num，若 group_num 过大可能导致溢出。
- 向量化变更需确保内存对齐和数据类型正确，可能引入隐蔽 bug。
- 专家缓存逻辑在网格迭代中未显式同步，潜在竞态条件风险。
- 优化仅基于特定测试配置，泛化性未验证，可能对其他参数产生回归。

影响：

- 用户：在兼容 GPU 上运行 MoE 模型时可能获得速度提升。
- 系统：修改核心 kernel，影响所有相关推理任务，需全面测试。
- 团队：提供优化范例，可推广至其他性能关键 kernel。

关联脉络

与近期 PR 关联：

- PR #6963 支持 NVFP4 MoE，共享 GPU 和量化优化主题。
- PR #7069 修复 MoE topk bug，涉及 MoE 层计算正确性。
- PR #7028 修复 KV 缓存量化，同样关注 GPU kernel 优化。

整体趋势显示团队持续投入 GPU kernel 性能优化和量化支持，本 PR 是该方向的具体实践。