PR #32616 完整报告

sgl-project/sglang

[JIT] Restore the previous division behavior in per-token group quantization

合并时间：2026-07-28 17:56

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/32616

执行摘要

本 PR 通过恢复单行除法操作，修复了 PR #30924 引入的 FP32 量化缩放精度退化，该退化导致 GLM-5.2 在 Hopper GPU 上 decode 性能大幅回落。变更极小（仅 1 行），但经过位级验证和性能测试，已确认修复效果。

功能与动机

Issue #32582 报告 GLM-5.2 decode 性能从 30 token/s 降至 10 token/s。通过二分法定位到 commit 4f45d011（来自 PR #30924），该提交在 per_token_group_quant.cuh 中将除法 kMaxValue / amax 改为显式倒数乘法 kMaxValue * __frcp_rn(amax)，在 Hopper 架构上引入微小 FP8 舍入差异，导致量化精度变化并拖慢解码。本 PR 恢复原除法，使 per-token-group 量化行为匹配之前 v2 内核。

实现拆解

1. 定位回归点：在 python/sglang/kernels/jit/csrc/gemm/per_token_group_quant.cuh 中，PR #30924 修改了 QuantTrait 结构体中 FP32 缩放值的计算方式。
2. 恢复除法：将第 308 行 kMaxValue * __frcp_rn(amax) 改回 kMaxValue / amax，消除倒数乘法的舍入偏差。
3. 验证：
   - H20 上 8 项 FP32 量化 CUDA 测试全部通过。
   - 与之前 v2 kernel 进行位级对比，FP8 code 和 scale 零差异。
   - 在 DeepSeek-V3.2 MTP 上，L3 接受长度从 2.462 恢复至 3.368（冷启动基线 3.459）。

python/sglang/kernels/jit/csrc/gemm/per_token_group_quant.cuh

唯一修改的文件，恢复除法操作以修复 FP32 量化缩放精度回归。

```
// python/sglang/kernels/jit/csrc/gemm/per_token_group_quant.cuh
// QuantTrait 结构体内的 per-token-group 量化缩放计算
// 第 305-308 行：FP32 scale 分支
} else {
    // fp32 scale: multiply in fp32 (hmul2 brings too much precision loss)
    scale_inv = raw_scale;
    // 恢复直接除法以避免 __frcp_rn 引入的舍入差异
    // 参见 PR #32582 和 #32616
    const float quant_scale = kMaxValue / amax;
```

```
    const float2 quant_scale2 = {quant_scale, quant_scale};
    #pragma unroll
    for (uint32_t i = 0; i < kVecSize / 2; ++i) {
      ...
    }
}
```

评论区精华

无 review 评论。作者在 PR body 中提供了详细的动机和验证结果，并获得 reviewer DarkSharpness 批准。

风险与影响

- 技术风险：低。单行恢复，位级一致性和性能测试已验证。但作者提及“exact mechanism is still unclear”，极端数值条件下可能存在未覆盖的舍入问题。
- 影响：直接影响 Hopper GPU 上使用 per-token-group 量化的模型（如 GLM-5.2）的 decode 性能。预期恢复到 PR #30924 之前的水平。不影响非量化路径或其他架构。

关联脉络

本 PR 与以下变更紧密相关：

- PR #30924：引入回归的原始变更，使用倒数乘法替代除法。
- Issue #32582：报告 GLM-5.2 decode 性能大幅下降的 bug report。
- PR #30393：在验证修复时使用的一个关联 PR，帮助恢复接受长度。

这些 PR 共同勾勒出一条围绕量化数值精度优化的演进线，本 PR 属于快速止损类修复，更根本的精度分析可能需要后续跟进。