PR #52775 完整报告

vllm-project/vllm

[Kernel] SM120: stop routing misaligned-M blockwise FP8 GEMMs to the small-M swapAB config

合并时间：2026-08-19 21:25

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/52775

执行摘要

本 PR 修复了 NVIDIA SM120 平台上 blockwise FP8 GEMM 的分发逻辑：将原先 M % 4 != 0 一律路由到 swapAB 小 M 配置的条件，收紧为仅当 M <= 64 时启用。该修复使 prefill 规模（大 M）的 misaligned GEMM 走回默认高性能路径，在 Qwen3.8-27B-FP8 模型上实现输出吞吐 +17.7%、TTFT 降低 28%、TPOT 降低 13% 的显著收益，且输出 bit-identical，无正确性风险。

功能与动机

PR body 明确指出：SM120 blockwise FP8 GEMM 分发将每个 M % 4 != 0 调用都路由到 swapAB 配置（TILE_N=32 ），而该配置是 #38325 为小 M/decode 场景添加的优化。在 V1 chunked prefill 下，混合批次常使 M 在 prefill 规模（M~8k）下未对齐，导致 82% 的 prefill 规模 blockwise GEMM 实例走了 swapAB 路径，耗时从 3.10ms 增至 4.84ms，端到端 prefill GEMM 时间约翻倍。这是性能回归，并非正确性问题。

实现拆解

1. 变更入口：csrc/libtorch_stable/quantization/w8a8/cutlass/c3x/scaled_mm_blockwise_sm120_fp8_dispatch.cuh 文件的 cutlass_gemm_blockwise_sm120_fp8_dispatch 函数中，swap_ab 的判定逻辑是唯一改动点。
2. 核心变更：将 bool swap_ab = (M <= 64) || (M % 4 != 0); 改为 bool swap_ab = (M <= 64);。这移除了 M % 4 != 0 对大 M 的误路由，使 misaligned 的大 M 调用落到默认 128x128x128 配置。
3. 配套验证：PR 通过微基准（M x N x alignment 矩阵）和端到端服务基准（并发 32、8K/1K token）验证了性能提升与输出 bit-identical。由于改动极小，未引入自动化测试，属于风险可控。

csrc/libtorch_stable/quantization/w8a8/cutlass/c3x/scaled_mm_blockwise_sm120_fp8_dispatch.cuh

核心修改文件：改变了 SM120 blockwise FP8 GEMM 的 swapAB 路由条件，直接影响 prefill 性能。

```
// SM120 blockwise FP8 GEMM 分发函数（节选）
// 原逻辑：swapAB 用于小 M 优化，但错误地包含了所有 M%4 != 0 的调用，
// 导致 prefill 规模（大 M）的 misaligned GEMM 走了次优路径。
// 修复后仅在小 M（decode 场景）启用 swapAB。
```

```
int M = a.size(0); // 获取 M 维度
// 更多启发式调优可在此根据 N/K 维度进行
bool swap_ab = (M <= 64); // 仅小 M 使用 swapAB 配置
if (!swap_ab) {
    if (M <= 256) {
        // 其他分支逻辑保持不变
    }
}
```

评论区精华

- claude[bot]：由于是 fork 的 PR，自动审查被禁用，建议维护者手动触发 @claude review。
- ZJY0516 直接批准（两次 APPROVED），未留下技术评论。

风险与影响

- 风险：核心 GEMM 分发路径变更，可能影响 SM120 上所有使用该函数的模型。但 PR 验证了 misaligned 点输出 bit-identical，且微基准覆盖了宽范围 M/N/K 组合。
- 影响：仅限于 NVIDIA SM120 平台。V1 引擎 + chunked prefill 场景下，prefill 性能显著提升；decode 场景无影响。团队可在 Blackwell 系列 GPU 上获得收益，但 SM90/SM100 仍保留原逻辑，未来可考虑迁移。

关联脉络

本 PR 直接修复了 #38325 引入的 swapAB 配置的适用条件。该 PR 属于 kernel 性能优化系列，与近期 SM120/Blackwell 相关的性能调优工作（如 FlashMLA sparse、sparse MLA mask 向量化）同属面向新硬件的性能改进方向。未来可能扩展至 SM90/SM100 的同类 dispatch 文件。