PR #52502 完整报告

vllm-project/vllm

[Hardware][NVIDIA] Add GB10 fused-MoE fp8 tuning configs (E=256, E=512)

合并时间：2026-08-17 11:04

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/52502

执行摘要

PR #52502 为 NVIDIA GB10（DGX Spark）新增两张 fused-MoE Triton 内核调优配置表，分别覆盖 DeepSeek 系 MoE 层的 E=256/N=512 与 E=512/N=512 形状（fp8_w8a8、block_shape=[128,128]）。改动为纯 JSON 配置新增（+294 行、0 删除），不涉及任何运行时代码路径；在 DGX Spark 真机 sweep 与 2× DGX Spark、TP=2 端到端验证中，MoE 层吞吐与 decode 吞吐相对启发式回退基线均有提升。PR 由 jeejeelee 批准、WoosukKwon 合并，与 #52192、#45949 共同构成 GB10 fused-MoE 配置族。

功能与动机

PR body 明确说明了动机："Without device-specific configs, GB10 falls back to heuristic defaults"——GB10 上 DeepSeek 系 MoE 层（fp8_w8a8、block_shape=[128,128]、N=512）此前没有设备特定配置，只能走启发式默认参数，性能未达最优。作者也预先回应了重复性问题：现有开放配置 #52192 是 E=512,N=2688（Nemotron-3-Super）、#45949 是 Qwen3-Coder-Next 形状，"No open or merged config covers device_name=NVIDIA_GB10 with N=512"。因此本 PR 补齐的是 E=256/512 且 N=512 这一具体形状缺口。

实现拆解

1. 变更入口与命名约定：在 vllm/model_executor/layers/fused_moe/configs/ 目录下新增两个 JSON 文件。文件名本身就是匹配键，编码了 E（专家数）、N（隐藏维）、device_name、dtype、block_shape 五个维度；fused-MoE 内核加载器在运行时按这些维度精确查找配置，命中后打印日志并采用参数，未命中则回退启发式默认。
2. 配置结构与分桶策略：每个文件顶层以 triton_version: "3.5.0" 作为版本锚点，避免跨 Triton 版本盲目套用；其下按 M（单次内核调用的 token 行数）划分 17 档（1 到 4096）。小 M 档（1-32）统一使用 BLOCK_SIZE_M=16、GROUP_SIZE_M=1，以细粒度分块降低小 batch 的跨行同步开销；M 进入 48 后 GROUP_SIZE_M 提升到 16，M>=256 再提升到 32，通过跨 token 行分组提升 A 矩阵访存复用；M>=1024 进入 prefill 长序列区间，BLOCK_SIZE_M 逐步升到 32/64，提高每个线程块的计算密度。BLOCK_SIZE_N 与 BLOCK_SIZE_K 全程固定 128，对应 block_shape=[128,128]；num_warps 固定为 4，num_stages 在 3/4 间切换，权衡流水线深度与寄存器占用。
3. 调优与验证流程：作者用 benchmarks/kernels/benchmark_moe.py 在 DGX Spark（GB10）上对每个 M 档 sweep 参数，随后在 2× DGX Spark、TP=2 的 DeepSeek-V4-Flash-0731 服务上做端到端验证：确认运行日志中两张配置均被选中、服务

稳定，且 decode 吞吐相对回退基线有提升。

4. 测试与配套：本 PR 没有任何源码、测试或文档改动；CI 由维护者 jeejeelee 的 /ci run 触发（Buildkite #84081）并通过后合并。PR body 声明调优由提交者在 GB10 硬件上完成，AI 仅用于把基于 v0.26.0 的生产分支 rebase 到 main。

vllm/model_executor/layers/fused_moe/configs/E=256,N=512,device_name=NVIDIA_GB10,dtype=fp8_w8a8,block_shape=[128,128].json

E=256 专家规模的 DeepSeek 系 MoE 层调优配置，是本 PR 的核心新增之一；文件名即运行时匹配键，编码 E、N、device_name、dtype、block_shape 五个维度，命中后由 fused-MoE 内核加载器采用。

```
{
    // 版本锚点：fused-MoE 内核加载器只在运行时的 Triton 主版本
    // 与此值一致时才加载本配置，避免跨版本盲目套用导致性能回退
    "triton_version": "3.5.0",

    // 分桶键为 M（单次内核调用的 token 行数），覆盖 1 ~ 4096，
    // 每个桶给出该 M 区间实测最优的 Triton launch 参数
    "1": {
        "BLOCK_SIZE_M": 16, // 小 batch 用 16，块粒度细、无效算力少
        "BLOCK_SIZE_N": 128, // 固定为 block_shape 的 N 维 128
        "BLOCK_SIZE_K": 128, // 固定为 block_shape 的 K 维 128
        "GROUP_SIZE_M": 1, // M 很小时按单行分组，避免跨行同步开销
        "num_warps": 4,
        "num_stages": 4 // 流水线深一点，容忍小 M 下的访存延迟
    },
    // 2/4/8/16/24 与 "1" 形状相同，仅 num_stages 降为 3，
    // 用更浅的流水线换取更低寄存器占用（此处省略重复条目）
    "32": {
        "BLOCK_SIZE_M": 16,
        "BLOCK_SIZE_N": 128,
        "BLOCK_SIZE_K": 128,
        "GROUP_SIZE_M": 1,
        "num_warps": 4,
        "num_stages": 4
    },
    // M 进入 48 以后 GROUP_SIZE_M 升为 16：跨多行 token 分组共享
    // A 矩阵分块，提升访存复用率，是中型 batch 的关键收益项
    "48": {
        "BLOCK_SIZE_M": 16,
        "BLOCK_SIZE_N": 128,
        "BLOCK_SIZE_K": 128,
        "GROUP_SIZE_M": 16,
        "num_warps": 4,
        "num_stages": 4
    },
    // 48 ~ 128 区间 GROUP_SIZE_M 统一为 16，num_stages 在 3/4 间微调；
    // 256 起 GROUP_SIZE_M 升到 32，继续放大跨 token 分组
```

```
    "256": {
        "BLOCK_SIZE_M": 16,
        "BLOCK_SIZE_N": 128,
        "BLOCK_SIZE_K": 128,
        "GROUP_SIZE_M": 32,
        "num_warps": 4,
        "num_stages": 3
    },
    // 1024 起进入 prefill 长序列区间，BLOCK_SIZE_M 逐步升到 32 / 64：
    // 用更大的 M 块提高每个线程块的计算密度
    "1024": {
        "BLOCK_SIZE_M": 32,
        "BLOCK_SIZE_N": 128,
        "BLOCK_SIZE_K": 128,
        "GROUP_SIZE_M": 16,
        "num_warps": 4,
        "num_stages": 3
    },
    "4096": {
        "BLOCK_SIZE_M": 64, // 4096 行时块内并行度充足，64 为最优
        "BLOCK_SIZE_N": 128,
        "BLOCK_SIZE_K": 128,
        "GROUP_SIZE_M": 32,
        "num_warps": 4,
        "num_stages": 4
    }
    // 1536 / 2048 / 3072 与 4096 形状相同（BLOCK_SIZE_M=64、GROUP_SIZE_M=32、
    // num_stages=4），此处省略重复条目
}
```

评论区精华

- claude[bot]：该 PR 来自 fork，自动审查默认关闭，维护者可通过 @claude review 触发一次性审查——实际审查中并未调用。
- jeejeelee：直接批准（无评论），并通过 /ci run 触发 Buildkite CI #84081。
- 值得注意：真正的 " 讨论 " 发生在 PR body 而非评论区——作者预先澄清了与 #52192、#45949 的重复关系，这是此类纯配置 PR 最容易被质疑的点。

风险与影响

- 兼容性风险低：纯 JSON 新增，配置查找机制为现成逻辑，其他设备的代码路径完全不受影响。
- 静默回退风险：若未来 Triton 升级导致 triton_version=3.5.0 不再命中，或 fused-MoE 内核实现 / 加载器解析方式变化，GB10 会静默回退到启发式默认，表现为性能回落而非报错，问题难以察觉。
- 覆盖范围有限：两张表只覆盖 N=512、fp8_w8a8、block_shape=[128,128] 的 DeepSeek 系形状；同一设备其他形状（如 #52192 的 N=2688）仍需各自的调优配置。

- 无自动化回归测试：配置有效性依赖提交者的一次性 sweep 与 E2E 验证，后续若内核分组逻辑、算子融合策略变化，这些参数可能不再最优。

关联脉络

- 52192：PR body 明确引用的现有 GB10 配置 PR，覆盖 E=512,N=2688（Nemotron-3-Super），与本次 N=512 形状互补，共同填充 GB10 fused-MoE 配置矩阵。
- 45949：PR body 引用的另一个 GB10 配置 PR（Qwen3-Coder-Next 形状），同属 GB10 enablement 演进线。
- 从同仓库近期 PR 看，vLLM 正在持续为 DeepSeek 系 MoE 做性能打磨（参见 #52212、#51967、#52084、#51538 等 DSV4 内核优化），本 PR 属于 " 按设备 + 形状 + 量化方式沉淀 tuning 配置 " 这一更广的配置治理方向，为 GB10 这类新硬件补齐了开箱即用的性能基线。