PR #44220 完整报告

vllm-project/vllm

[Perf] use triton moe backend on hopper by default

合并时间：2026-06-02 15:52

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/44220

执行摘要

- 一句话：Hopper 上默认使用 Triton MoE 后端
- 推荐动作：建议合并。该 PR 基于实际基准测试数据，将 Hopper 上 MoE 后端的默认选择从 FlashInfer 切换为 Triton，性能提升明确，风险低。值得关注的是 Hopper 特定优化和基准测试方法，可推广到类似决策中。

功能与动机

根据 Issue #41306 报告，v0.20 版本中 MoE 模型（如 Mixtral-8x7B）在 Hopper 架构上出现了显著的延迟和吞吐量回归。作者在 H200 上使用 Qwen3-30B-A3B 模型进行基准测试，发现 Triton 后端的性能优于 FlashInfer Cutlass 后端（吞吐量 57821 vs 55700 tokens/s），因此建议在 Hopper 上默认使用 Triton 后端。

实现拆解

该 PR 仅修改了一个文件 vllm/model_executor/layers/fused_moe/oracle/unquantized.py，在 _get_priority_backends 函数中为 CUDA 平台添加了 Hopper (SM90) 特定的后端优先级调整：

1. 检测 SM90 架构：使用 current_platform.is_device_capability_family(90) 判断是否为 Hopper 平台。
2. 降低 FlashInfer 优先级：通过 _move_to_back 函数将 FLASHINFER_TRTLLM 和 FLASHINFER_CUTLASS 从列表头部移到尾部，使得列表中后续的 TRITON 和 BATCHED_TRITON 成为首选。
3. 保持现有逻辑：原有的 DP 大小大于 1 时的 FLASHINFER_CUTLASS 降级逻辑仍然保留，且 Hopper 特殊处理在该逻辑之前执行，避免冲突。

该变更不涉及配置、测试或部署改动，所有更改都集中在后端选择器的控制流中。

关键文件：

- vllm/model_executor/layers/fused_moe/oracle/unquantized.py（模块 MoE 后端；类别 source；类型 data-contract；符号 _get_priority_backends）：核心变更文件，修改了 MoE 后端优先级选择逻辑，新增对 Hopper (SM90) 架构的特殊处理。

关键符号：_get_priority_backends

关键源码片段

vllm/model_executor/layers/fused_moe/oracle/unquantized.py

核心变更文件，修改了 MoE 后端优先级选择逻辑，新增对 Hopper (SM90) 架构的特殊处理。

```
# vllm/model_executor/layers/fused_moe/oracle/unquantized.py

elif current_platform.is_cuda():
    _AVAILABLE_BACKENDS = [
        UnquantizedMoeBackend.FLASHINFER_TRTLLM, # 原本是最高优先级
        UnquantizedMoeBackend.FLASHINFER_CUTLASS, # 原本是第二优先级
        UnquantizedMoeBackend.TRITON, # 原本是第三优先级
        UnquantizedMoeBackend.BATCHED_TRITON, # 原本是第四优先级
    ]

    # On Hopper (SM90), the FlashInfer unquantized MoE kernels are slower
    # than Triton, so prefer Triton by default.
    if current_platform.is_device_capability_family(90):
        # 将 FlashInfer 后端移到列表末尾，使 Triton 成为首选
        _move_to_back(_AVAILABLE_BACKENDS,
                      UnquantizedMoeBackend.FLASHINFER_TRTLLM)
        _move_to_back(_AVAILABLE_BACKENDS,
                      UnquantizedMoeBackend.FLASHINFER_CUTLASS)

    # HACK: Qwen3.5 has crash with FLASHINFER_CUTLASS BF16 if DEP.
    # Updating the oracle querying logic is out of the scope of this
    # PR. Need to fix the kernel or update structure in follow up.
    if moe_config.moe_parallel_config.dp_size > 1:
        _move_to_back(_AVAILABLE_BACKENDS,
                      UnquantizedMoeBackend.FLASHINFER_CUTLASS)
```

评论区精华

审阅者 mgoin 对性能测试的全面性提出了建议，认为在 Expert Parallelism（EP）或特定规模的 MoE 下，FlashInfer 可能仍然有优势，建议进行更多基准测试。作者 ZJY0516 回应称更多基准结果已附在关联 Issue #41306 中。该讨论未被解决，但 PR 仍被另一审阅者 zyongye 批准。

- Hopper 上 FlashInfer 后端性能是否在 EP 场景下更好 (performance): 作者提供了基准测试链接，但未做进一步测试。未达成明确结论，PR 仍通过。

风险与影响

- 风险：风险较低：
  - 仅修改了后端选择逻辑，且仅在 SM90 平台上生效，不影响其他 GPU 架构（如 Ampere、Ada Lovelace）或其他平台（ROCm、XPU、CPU）。
  - 修改的是优先级顺序而非删除后端，用户仍可通过环境变量或配置显式指定 FlashInfer 后端。

- 潜在风险：如果某些 MoE 模型在 Hopper 上对 Triton 后端存在兼容性问题，可能会导致运行时错误。但根据现有测试，Triton 后端已经过广泛使用。
- 影响：对 Hopper 架构（如 H100、H200）用户有正面性能影响，预计 MoE 模型吞吐量提升约 4%。对非 Hopper 平台无影响。代码维护成本极低，仅增加了 6 行条件逻辑。
- 风险标记：缺少更全面的基准测试验证，仅针对 Hopper 架构，需确认其他 SM90 子型号兼容性

关联脉络

- PR #41306 [Bug]: v0.20 latency and throughput regression on MoE models: 直接关联的 Issue，报告了 MoE 模型在 v0.20 的性能回归，本 PR 旨在缓解该问题。