PR #29470 完整报告

sgl-project/sglang

[GLM-5] Tune the threshold of router GEMM

合并时间：2026-06-30 05:19

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/29470

执行摘要

- 一句话：调优 GLM-5 MoE 路由 GEMM 阈值以提升性能
- 推荐动作：该 PR 值得精读，尤其适合关注 GPU kernel 调度和性能调优的工程师。其展示了如何通过微基准测试和 E2E 吞吐对比，平衡 micro-benchmark 与实际负载的差异。设计决策（架构感知阈值）简洁有效。

功能与动机

作者指出原全局阈值 16 在高并发（如 8-16 tokens）下，PyTorch 的 splitK GEMM 实现无法有效与后续 kernel 进行 PDL 重叠，导致 E2E 性能劣化。通过微基准测试和实际吞吐对比，确定 SM10X 平台更适合 4 token 的阈值。

实现拆解

1. 调整阈值条件：在 python/sglang/srt/models/deepseek_v2.py 的 FusedMoE 类的 forward 方法中，将原来硬编码的 hidden_states.shape[0] <= 16 改为动态阈值。
2. 引入架构感知阈值：新增变量 max_router_gemm_tokens，对于 SM100 或 SM103（Blackwell 架构）设为 4，其余架构（SM>=90）沿用 16。
3. 代码注释：添加说明注释 # NOTE(b8zhong): this threshold has been empirically verified，表明阈值经过实证验证。
4. 无测试改动：该 PR 未修改任何测试文件。

关键文件：

- python/sglang/srt/models/deepseek_v2.py（模块 模型；类别 source；类型 data-contract；符号 FusedMoE.forward）：修改了 MoE 路由 GEMM 的触发阈值，影响 GLM-5 等 DeepSeek 模型的解码性能。

关键符号：FusedMoE.forward

关键源码片段

python/sglang/srt/models/deepseek_v2.py

修改了 MoE 路由 GEMM 的触发阈值，影响 GLM-5 等 DeepSeek 模型的解码性能。

```
# python/sglang/srt/models/deepseek_v2.py # 在 FusedMoE.forward 中调整了 JIT 路由 GEMM 的触发阈值
```

```
# 原来的硬编码 16 被替换为架构感知的 max_router_gemm_tokens
# 新阈值：对于 SM100 和 SM103 (Blackwell) 设为 4，其他 SM>=90 保持 16

# NOTE(b8zhong): this threshold has been empirically verified
# 通过 1K/1K 交叉测试 (token 1-16 sweep) 确定 SM10X 上 4 token 阈值最佳
# 目的是避免高并发下 PyTorch splitK 无法与后续 kernel 进行 PDL 重叠
max_router_gemm_tokens = 4 if _device_sm in (100, 103) else 16
if (
    _is_cuda
    and hidden_states.shape[0] <= max_router_gemm_tokens # 此处原本为 <= 16
    and hidden_states.shape[1] % 1024 == 0
    and (self.weight.shape[0] == 256 or self.weight.shape[0] == 384) # 只影响 MoE 路由 GEMM (
    权重行数 256 或 384)
    and _device_sm >= 90
):
    logits = _jit_dsv3_router_gemm(
        hidden_states, self.weight, out_dtype=torch.float32
    )
```

评论区精华

该 PR 的 review 讨论较少。Fridge003 给予了批准，无额外评论。作者在 PR body 中提供了详细的性能数据表格，对比了不同并发下的延迟和吞吐，并与 PR#21531（引入该 kernel）和 vllm PR#44217、lightseek PR#33 进行了关联讨论。

- 暂无高价值评论线程

风险与影响

- 风险：
  1. 回归风险低：仅在特定架构（SM100/SM103）上降低阈值，且只影响路由 GEMM 分支，其他架构行为不变。性能数据表明无退化。
  2. 影响范围有限：仅修改一个分支条件，无 API 或数据结构变更。
  3. 缺少动态自适应：阈值仍为静态值，可能随 PyTorch 版本或模型变化需要重新调优。
- 影响：
  - 用户：GLM-5.2 用户在 SM10X 显卡（如 B200）上可获得更稳定的解码性能，低并发吞吐提升约 1-2%，高并发持平或微升。
  - 系统：无新增依赖或配置项，对非 DeepSeek 模型无影响。
  - 团队：降低后续微基准测试中路由 GEMM 的干扰，为未来类似调整提供了方法论参考（交叉测试 + 架构感知阈值）。
  - 风险标记：核心路径变更

关联脉络

- PR #21531 [Model] Add DeepSeek V3.2 model (JIT router GEMM): 引入了被调优的路由 GEMM kernel，是该 PR 的前置依赖。