PR #25509 完整报告

sgl-project/sglang

[misc] Throw error when single batch overlap is enabled on Hopper

合并时间：2026-05-19 05:51

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25509

执行摘要

- 一句话：Hopper GPU 上禁用 SBO 特性
- 推荐动作：PR 改动小且明确，建议合并。

功能与动机

PR #25491 将后续重新支持该特性，当前在 Hopper GPU 上启用 SBO 会导致问题，因此需要提前报错引导用户。

实现拆解

1. 在 python/sglang/srt/layers/moe/utils.py 的 initialize_moe_config 函数中，在 IS_SBO_ENABLED 赋值之后，添加运行时检查：如果 IS_SBO_ENABLED 为真且 CUDA 可用，则通过 torch.cuda.get_device_capability()[0] == 9 检测是否为 SM90（Hopper）架构。
2. 若条件成立，抛出 ValueError，提示用户移除 --enable-single-batch-overlap 参数。

关键文件：

- python/sglang/srt/layers/moe/utils.py（模块 MoE 配置；类别 source；类型 core-logic）：核心文件，在 initialize_moe_config 中添加了 SBO 在 Hopper 上的运行时检查。

关键符号：initialize_moe_config

关键源码片段

python/sglang/srt/layers/moe/utils.py

核心文件，在 initialize_moe_config 中添加了 SBO 在 Hopper 上的运行时检查。

```
# python/sglang/srt/layers/moe/utils.py

def initialize_moe_config(server_args: ServerArgs):
    global MOE_A2A_BACKEND, MOE_RUNNER_BACKEND
    global SPECULATIVE_MOE_RUNNER_BACKEND, SPECULATIVE_MOE_A2A_BACKEND
    global DEEPEP_MODE, DEEPEP_CONFIG
    global IS_TBO_ENABLED, IS_SBO_ENABLED
    global TBO_TOKEN_DISTRIBUTION_THRESHOLD
    global DISABLE_FLASHINFER_CUTLASS_MOE_FP4_ALLGATHER
```

```
    global MOE_QUANTIZATION

    # ... 其他初始化 ...

    IS_TBO_ENABLED = server_args.enable_two_batch_overlap
    IS_SBO_ENABLED = server_args.enable_single_batch_overlap

    # 新增：如果启用了 SBO 且 GPU 为 SM90（Hopper），则报错
    if IS_SBO_ENABLED and torch.cuda.is_available():
        if torch.cuda.get_device_capability()[0] == 9:
            raise ValueError(
                "SBO (single batch overlap) is not supported on SM90 GPUs "
                "with latest sgl-deep-gemm wheel. Please try removing "
                "--enable-single-batch-overlap argument."
            )

    TBO_TOKEN_DISTRIBUTION_THRESHOLD = server_args.tbo_token_distribution_threshold
    # ... 后续配置 ...
```

评论区精华

无 review 讨论。

- 暂无高价值评论线程

风险与影响

- 风险：变更仅 5 行新增代码，风险极低。仅影响 Hopper GPU 上使用 SBO 的用户，会提前失败并给出清晰错误信息。
- 影响：用户影响：在 Hopper GPU 上使用 --enable-single-batch-overlap 的用户将遇到 ValueError 启动时错误，需移除该参数。系统影响：不会影响其他 GPU 架构或不使用 SBO 的场景。
- 风险标记：启动时错误抛出

关联脉络

- PR #25491 [misc] 相关 Issue: PR 正文提及关联 issue #25491，该 issue 计划后续重新支持 SBO 特性。