PR #28749 完整报告

sgl-project/sglang

Fix nightly CI test for GLM-4.6 + B200

合并时间：2026-06-25 03:34

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/28749

执行摘要

- 一句话：修复 GLM-4.6 在 B200 上的 CI 失败
- 推荐动作：建议快速合入。这是一个针对硬件特定配置的精准修复，改动量小且经过验证。

功能与动机

Nightly CI test for GLM-4.6 on B200 始终失败，而本地测试正常。PR body 提到可能权重下载不完整，但实际根因是 B200 硬件（SM100）上的 MoE runner 后端未正确配置。

实现拆解

1. Python/sglang/srt/server_args.py: 在 Glm4MoeForCausalLM 模型分支中，将条件判断从 self.quantization == "modelopt_fp4" 扩展为 self.quantization in {"modelopt_fp4", None}。这样当量化参数为 None（即未量化）时，也能设置 moe_runner_backend = "flashinfer_trtllm"，确保 SM100 上使用正确的 MoE 后端。
2. scripts/ci/utils/slash_command_handler.py: 修改 _LEGACY_SUITE_TO_RUNNER_CONFIG 映射，将 "nightly-8-gpu-common" 的值从单个字符串 "8-gpu-h200" 改为列表 ["8-gpu-h200", "8-gpu-b200"]。同时更新 detect_suite 函数以处理列表值，使 /rerun-test 命令能同时将测试分发到 H200 和 B200。

关键文件：

- python/sglang/srt/server_args.py（模块 模型配置；类别 source；类型 core-logic）：核心修复：放宽 Glm4MoeForCausalLM 模型在 SM100 上 MoE runner 后端的条件判断，支持未量化场景。
- scripts/ci/utils/slash_command_handler.py（模块 CI 脚本；类别 infra；类型 infrastructure）：修复 /rerun-test 命令对 nightly-8-gpu-common 套件的分发逻辑，确保同时运行在 H200 和 B200。

关键符号：未识别

关键源码片段

python/sglang/srt/server_args.py

核心修复：放宽 Glm4MoeForCausalLM 模型在 SM100 上 MoE runner 后端的条件判断，支持未量化场景。

```
# python/sglang/srt/server_args.py ( 第 4329-4337 行 )
            if (
                self.quantization in {"modelopt_fp4", None} # 原为 == "modelopt_fp4"，现支持
                None（未量化）
                and self.moe_a2a_backend == "none"
                and self.moe_runner_backend == "auto"
            ):
                self.moe_runner_backend = "flashinfer_trtllm"
                logger.info(
                    "Use flashinfer_trtllm as MoE runner backend on sm100 for
                    Glm4MoeForCausalLM"
                )
```

scripts/ci/utils/slash_command_handler.py

修复 /rerun-test 命令对 nightly-8-gpu-common 套件的分发逻辑，确保同时运行在 H200 和 B200。

```
# scripts/ci/utils/slash_command_handler.py

# 第 728 行：映射改为列表，同时分发到 H200 和 B200
"nightly-8-gpu-common": ["8-gpu-h200", "8-gpu-b200"],

# detect_suite 函数（第 833-843 行）适配列表值
if mappable:
    results = []
    for s in mappable:
        rcs = _LEGACY_SUITE_TO_RUNNER_CONFIG[s]
        if isinstance(rcs, str):
            rcs = [rcs] # 兼容旧格式
        for rc in rcs:
            results.append(_resolve_runner_config(rc, full_path, s))
    return results
```

评论区精华

无人工 review 讨论。仅 Gemini 代码助手自动评论确认变更直观。

- 暂无高价值评论线程

风险与影响

- 风险：风险较低。server_args.py 的改动仅放宽条件判断，加入 None 值，不影响现有量化模型的逻辑。CI 脚本改动将 nightly-8-gpu-common 套件额外分发到 B200，可能增加 CI 资源消耗，但属于预期行为。
- 影响：直接影响：GLM-4.6 模型在 B200（SM100）上能正确配置 MoE runner，修复 nightly CI 测试。间接影响：/rerun-test 命令现在会将 nightly-8-gpu-common 套件同时分发到 H200 和 B200，确保 Blackwell 硬件上的测试覆盖。
- 风险标记：核心路径变更 , CI 配置变更

关联脉络

- PR #28144 Fix TRTLLM MHA FP8 KV *cache scale* handling: 同为 B200（Blackwell）硬件相关的 bugfix，涉及 SM100 上的注意力或 MoE 后端配置。