

PR #25453 完整报告

sgl-project/sglang

[CI] Lower mem-fraction-static for GLM-5.1 FP8 8-GPU test to 0.85

合并时间: 2026-05-16 11:14

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/25453>

执行摘要

- 一句话: 降低 GLM-5.1 FP8 测试内存比例至 0.85 修复 B200 OOM
- 推荐动作: 该 PR 为一次小范围测试配置修复, 无精读必要。但可关注其根因分析思路: B200 显存基线比 H200 高约 5 GiB, 导致同样参数下 OOM。对所有 DP-attention 测试的 mem-fraction-static 统一管理值得借鉴。

功能与动机

`test_glm51_fp8` 在 B200 夜班 CI 中持续 OOM 失败约 37 天 (自 #22399 引入), TP8+DP8 变体在 scheduler 初始化时因 `--mem-fraction-static=0.9` 导致内存不足 (178 GiB/GPU 下仅剩 ~3 GiB 空闲, 无法分配 6.38 GiB 的 cuda-graph)。而 TP8 和 TP8+DP8+MTP 变体正常。PR 作者通过本地 8xB200 复现确认 OOM 签名与 CI 一致。

实现拆解

1. 修改测试参数: 在 `test/registered/8-gpu-models/test_glm_51_fp8.py` 的 `COMMON_ARGS` 列表中将 `--mem-fraction-static` 从 0.9 改为 0.85。
2. 验证兼容性: 本地 8xB200 上测试 TP8、TP8+DP8、TP8+DP8+MTP 三种变体均通过, GSM8K 精度与 0.9 版本一致 (约 0.967)。
3. 对齐现有实践: 同一目录下其他 8-GPU + DP-attention 测试已使用 0.85 或更低值, 该变更保持了测试配置间的一致性。GB300 测试 (288 GiB/GPU) 维持 0.9。

关键文件:

- `test/registered/8-gpu-models/test_glm_51_fp8.py` (模块 测试; 类别 test; 类型 test-coverage): 核心变更文件, 将 `--mem-fraction-static` 从 0.9 改为 0.85, 直接修复 B200 OOM。

关键符号: 未识别

关键源码片段

`test/registered/8-gpu-models/test_glm_51_fp8.py`

核心变更文件, 将 `--mem-fraction-static` 从 0.9 改为 0.85, 直接修复 B200 OOM。

```
# test/registered/8-gpu-models/test_glm_51_fp8.py
import unittest
```

```
from sglang.test.accuracy_test_runner import AccuracyTestParams
from sglang.test.ci.ci_register import register_cuda_ci
from sglang.test.performance_test_runner import PerformanceTestParams
from sglang.test.run_combined_tests import run_combined_tests
from sglang.test.test_utils import ModelLaunchSettings

# Runs on both H200 and B200 via nightly-8-gpu-common suite
register_cuda_ci(est_time=1800, suite="nightly-8-gpu-common", nightly=True)

GLM_51_FP8_MODEL_PATH = "zai-org/GLM-5.1-FP8"

# 关键变更: mem-fraction-static 从 0.9 下调至 0.85, 避免 B200 上 TP8+DP8 OOM
COMMON_ARGS = [
    "--trust-remote-code",
    "--reasoning-parser=glm45",
    "--tool-call-parser=glm47",
    "--mem-fraction-static=0.85", # 原为 0.9, B200 上导致 scheduler 初始化 OOM
    "--enable-metrics",
]
```

评论区精华

无 review 讨论, 仅 bot 消息和 [/tag-and-rerun-ci](#) 命令。PR 获得 b8zhong APPROVED。

- 暂无高价值评论线程

风险与影响

- 风险:
 - 回归风险: 低。仅在测试配置层面微调一个参数, 精度验证显示无退化。
 - 性能风险: 低。mem-fraction-static 降低可能略微减少可用 KV cache 空间, 但测试中 GSM8K 精度无损失, 且非 TP8+DP8 变体性能同样稳定。
 - 兼容性: 仅影响 B200 测试, H200 和 GB300 不受影响。
- 影响:
 - 用户影响: 无直接影响, 变更仅涉及 CI 测试配置。
 - 系统影响: 修复 B200 夜班 CI 的持续失败, 提高 CI 可靠性。
 - 团队影响: 减少因测试不稳定导致的排查工作。
 - 风险标记: 暂无

关联脉络

- PR #22399 Add GLM-5.1 FP8 test: 该 PR 首次引入本测试文件, 原参数 0.9 导致 B200 OOM; 当前 PR 对其进行修正。