

PR #29502 完整报告

sgl-project/sglang

[CI] Fix GB300 DSV4 Pro FP4 nightly

合并时间: 2026-06-28 12:06

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/29502>

执行摘要

- 一句话: 修正 GB300 FP4 nightly 测试内存比例
- 推荐动作: 简单修复, 无需深入阅读。但可留意是否后续需要为不同变体动态调整内存比例, 避免硬编码。

功能与动机

修复 GB300 DSV4 Pro FP4 nightly CI 测试失败。PR body 未详细说明原因, 但从变更内容看, high-throughput 变体内存比例 0.85 导致 OOM 或性能不达标, 需调整至 0.9。

实现拆解

1. 修改 test/registered/gb300/test_deepseek_v4_pro_fp4.py 中 HIGH_THROUGHPUT_ARGS 列表的 --mem-fraction-static 参数值, 从 "0.85" 改为 "0.9"。
2. 提交历史显示作者曾临时禁用其他 GB300 nightly 变体以隔离问题, 最终恢复并仅保留此调整。

关键文件:

- test/registered/gb300/test_deepseek_v4_pro_fp4.py (模块 测试脚本; 类别 test; 类型 test-coverage): 唯一变更文件, 调整 high-throughput 变体的内存比例参数。

关键符号: test_deepseek_v4_pro_fp4

关键源码片段

`test/registered/gb300/test_deepseek_v4_pro_fp4.py`

唯一变更文件, 调整 high-throughput 变体的内存比例参数。

```
# 文件 : test/registered/gb300/test_deepseek_v4_pro_fp4.py
# 关键变更 : high-throughput 变体的内存比例从 0.85 调整为 0.9
```

```
HIGH_THROUGHPUT_ARGS = [
    "--trust-remote-code",
    "--dp",
    "4",
    "--enable-dp-attention",
    "--moe-a2a-backend",
```

```
"megamoe",
"--mem-fraction-static",
"0.9", # 之前为 "0.85", 提升以解决 nightly 测试失败
"--cuda-graph-max-bs",
"128",
"--max-running-requests",
"256",
]
```

评论区精华

Gemini Code Assist bot 提出评论：注释掉的 low-latency 和 balanced 变体会留下死代码，建议添加 TODO 说明或彻底删除。但该评论针对的是中间提交中临时禁用其他变体的操作，最终提交已恢复这些变体，因此该建议已过时。最终审核者 Fridge003 批准了 PR。

- 禁用变体残留代码清理 (style): 中间提交中临时禁用了其他变体，但最终提交已恢复，因此该建议不再适用。

风险与影响

- 风险：风险极低。仅修改测试配置文件中的一个数值，不影响任何生产逻辑。内存比例从 0.85 调整到 0.9，可能略微增加 GPU 内存使用，但对于 GB300 (4xB200) 而言在安全范围内。
- 影响：影响范围仅限于 GB300 上 DeepSeek-V4-Pro FP4 nightly CI 测试的 high-throughput 变体。其他变体和生产环境不受影响。团队可期待 nightly 测试更稳定。
- 风险标记：暂无

关联脉络

- 暂无明显关联 PR